

Université de Montréal

Prédiction et visualisation de la réactivité chimique des ARN

**Proposition d'un modèle basé sur la réactivité
des cycles minimaux et des sous-structures composées de ces
cycles.**

par Philippe Malric

Département de biochimie
Faculté de Médecine

Mémoire présenté
en vue de l'obtention du grade de
Maître en Bio-informatique

Octobre, 2018

© Philippe Malric, 2018

Résumé

Mesurer la réactivité chimique d'un nucléotide d'une séquence d'ARN permet d'inférer sa conformation locale. L'interprétation la plus simple de la conformation locale d'un nucléotide est de le qualifier par le terme pairé ou non pairé. Bien qu'ils soient reliés à la réactivité chimique des nucléotides, ces deux états ne l'expliquent pas entièrement.

Actuellement, la « *RNA mapping database* » (RMDB) contient 136 892 séquences sondées dans différentes conditions, ce qui donne un bon jeu de données pour étudier la structure des ARN en utilisant des algorithmes d'apprentissages machines. Dans mon projet de recherche, j'ai tenté de prédire et de comprendre la réactivité ou la non-réactivité des nucléotides à partir de plusieurs ensembles de données obtenues de la RMDB. Pour cela, j'ai utilisé deux algorithmes de repliement des ARN en structures secondaires : MCFlashFold du laboratoire de Dr François Major et RNAsubopt de la collection Vienna package. L'originalité de mon projet de recherche est qu'il se base sur une abstraction des composants de la structure secondaire, les cycles et les sous-structures, pour caractériser les nucléotides.

Dans ce mémoire, je démontre que la majorité des valeurs de réactivités discrètes (hautes/basses) des nucléotides peuvent être prédites, bien que certaines valeurs soient surprenantes. En effet, dans l'ensemble de données testé, il y a plus de nucléotides peu réactifs non pairés que de nucléotides réactifs non pairés. Il est donc légitime de se poser la question suivante : peut-on améliorer ces résultats? La réponse est oui et le modèle que j'ai utilisé pour le démontrer utilise des cycles minimaux au lieu des paires de bases. En titre d'illustration suivez : <http://majsrv1.irc.ca:3000/RDV>.

Mots-clés : ARN, sous-structure, structure secondaire, SHAPE, Eterna, RMDB

Abstract

Measuring the chemical reactivity of a nucleotide within a sequence allows to infer the local conformation of a nucleotide most of the time. The simplest interpretation of the local conformation of a nucleotide is by the terms paired or non-paired. However, these two states do not fully explain the reactivity of nucleotides in chemical probing experiments.

Currently, the RNA mapping database (RMDB) contains 136,892 sequences probed under different conditions, which provides us a relatively good data set to learn using machine learning algorithms. In my research project, I attempted to predict and understand the reactivity or non-reactivity of nucleotides of thousands of RNA obtained from RMDB. For this, I used two secondary structure RNA folding algorithms: MCFlashFold from Dr. François Major's laboratory and RNAsubopt from the Vienna package collection. The originality of my project is that it is based on two abstractions of the secondary structure called cycle and substructure.

In this thesis, I demonstrate that the majority of the discretized (high / low) reactivity values of nucleotides can be predicted, but not all. In fact, in the tested dataset, there are more predicted paired than predicted non-paired nucleotides that are reactive to 1m7. As I show here, this result can be improved by using minimum cycle instead of base pair.

The newly developed interface, named RDV, allows to quickly and easily compare homolog nucleotides between RNA. (See: <http://majsrv1.irc.ca:3000/RDV>).

Keywords: RNA, sub-structure, secondary structure, SHAPE, Eterna, RMDB

Table des matières

RÉSUMÉ	3
ABSTRACT	4
TABLE DES MATIÈRES	5
LISTE DES TABLEAUX	7
LISTE DES FIGURES	8
LISTE DES SIGLES	10
LISTE DES ABRÉVIATIONS	10
REMERCIEMENTS	12
AVANT-PROPOS	13
1.1 AVERTISSEMENT	13
1.2 MISE EN CONTEXTE	13
1.3 POURQUOI ÉTUDIER LA STRUCTURE DE L'ARN	16
INTRODUCTION.....	17
2.1 HISTORIQUE	18
2.2 DÉFINITION DES TERMES ET DES CONCEPTS UTILISÉS POUR PRÉDIRE LA RÉACTIVITÉ DES NUCLÉOTIDES D'UN ARN. 20	
CHAPITRE 1: RNASS_V2 : OBTENIR LA RÉACTIVITÉ DES S-S.....	33
3.1 DE L'OBTENTION DES DONNÉES À LA PRÉDICTION DISCRÈTE	33
3.2 CRÉATION DES ENSEMBLES DE DONNÉES.....	40
3.3 PRÉDICTIONS DE LA RÉACTIVITÉ DES NUCLÉOTIDES	48
CONCLUSION DU CHAPITRE 1	49
CHAPITRE 2 : RDV : VISUALISATION DE LA SS DES ARN	50
4.1 VISUALISATION DU GRAPHE DES TRANSITIONS.....	50
4.2 VISUALISATION DE LA COHÉRENCE	51
4.3 VISUALISATION DE LA SS	51
4.4 OBTENIR DES DÉTAILS ET RECHERCHER DES ARN SEMBLABLES	51

CONCLUSION DU CHAPITRE 2	54
CHAPITRE 3 : ÉVALUATION DU MODÈLE DES CYCLES SIMPLES	55
5.1 COMPARAISON DES PRÉDICTIONS DE RNASS AVEC CELLES FAITES À L'AIDE DE L'ÉTAT PAIRÉ OU NON D'UN NUCLÉOTIDE	57
CONCLUSION DU CHAPITRE 3	61
CONCLUSION	62
CHAMP D'ÉTUDE À VENIR ET ALGORITHMES À CONSIDÉRER	63
BIBLIOGRAPHIE	I
ANNEXE	I
CARACTÉRISTIQUES DU SERVEUR UTILISÉ	I
API D'ETERNA	I
ÉVALUATION DU SCORE DE PRÉDICTION.....	II
SCORE DE PRÉDICTIONS DE LA SS DE DE LA <i>MFE</i> DANS L'ENSEMBLE NON FILTRÉ.....	II
PORTRAIT D'UN ARN.....	VII

Liste des tableaux

TABLEAU I.	REPRÉSENTATION DE DEUX CYCLES SIMPLES	24
TABLEAU II.	EXEMPLE D'INFORMATION SUR LES EXPÉRIENCES DE LA RMDB. (ETERNA_R00_0000)	34
TABLEAU III.	CATÉGORIES DE NT. EN FONCTION DE LEUR RÉACTIVITÉ	43
TABLEAU IV.	PROPRIÉTÉ DU CARRÉ POUR DES VALEURS ENTRE 0 ET 1	44
TABLEAU V.	TABLEAU COMPARATIF DES PARAMÈTRES DE L'ENSEMBLE D'ENTRAÎNEMENT DU LOGICIEL RNASUBOPT	47
TABLEAU VI.	TABLEAU COMPARATIF DES PARAMÈTRES DE L'ENSEMBLE D'ENTRAÎNEMENT DU LOGICIEL MCFLASHFOLD.	47
TABLEAU VII.	TABLEAU DE CONTINGENCE DE L'ÉTAT PAIRÉ OU NON DES NT. EN FONCTION DE LEUR NIVEAU DE RÉACTIVITÉ.	57
TABLEAU VIII.	PERFORMANCE DES ALGORITHMES D'APPRENTISSAGE MACHINES À PRÉDIRE LA RÉACTIVITÉ CHIMIQUE DES NT..	60

Liste des figures

FIGURE 1. NOMBRE DE SÉQUENCES D'ARN AJOUTÉES À LA RMDB ENTRE SEPTEMBRE 2014 ET MARS 2017.	15
FIGURE 2. REPRÉSENTATION D'UN NUCLÉOTIDE, L'ADÉNOSINE MONOPHOSPHATE (AMP).....	20
FIGURE 3. REPRÉSENTATION EN BALLES ET BÂTONS (<i>BALLS AND STICKS</i>) D'UN GROUPEMENT PHOSPHATE ET D'UN BETA-D-RIBOSE.	21
FIGURE 4. LES QUATRE NUCLÉOTIDES DES ARN.....	22
FIGURE 5. EXEMPLE DE VISUALISATION D'UNE SS GÉNÉRÉE PAR LE LOGICIEL <i>RNA DYNAMIQUE VIEWER</i> . (VOIR CHAPITRE 2)	23
FIGURE 6. UN CYCLE « 3_2 ».	25
FIGURE 7. UNE S-S FORMÉE DE DEUX CYCLES ACCOLÉS.	25
FIGURE 8. REPRÉSENTATION DE TROIS CYCLES SE CHEVAUCHANT.	26
FIGURE 9. DIAGRAMME À RAIL DE L'EXPRESSION RÉGULIÈRE ASSOCIÉE À L'IDENTIFIANT DES S-S.....	26
FIGURE 10. RÉACTION CHIMIQUE DU 1M7 AVEC LE 2'OH DU SUCRE D'UN NUCLÉOTIDE.	32
FIGURE 11. CARTE DE CHALEUR DE L'EXPÉRIENCE : « ETERNA_R00_0000 ».	35
FIGURE 12. LES DONNÉES DE RÉACTIVITÉS SONT ÉGALEMENT PHASÉES POUR RDV ET LE SITE WEB DE LA RMDB.....	36
FIGURE 13. COMPARAISON ENTRE UNE STRUCTURE PRODUITE PAR <i>RNA DYNAMIC VIEWER</i> ET SON HOMOLOGUE PRIS DE LA PAGE WEB DE LA RMDB.	37
FIGURE 14. SS PROVENANT DE FORNA.	38
FIGURE 15. ILLUSTRATION DE L'ALGORITHME D'IDENTIFICATION DES S-S.	39
FIGURE 16. LA DISTRIBUTION DE LA DIVERSITÉ D'ENSEMBLE DE 1000 ARN PROVENANT DE LA RMDB.	41
FIGURE 17. LA DISTRIBUTION DE LA RÉACTIVITÉ MOYENNE DE 2500 ARN PROVENANT DE LA RMDB.....	42
FIGURE 18. LA DISTRIBUTION DU « <i>SIGNAL TO NOISE</i> » DE 2500 ARN PROVENANT DE LA RMDB.	42
FIGURE 19. COURBE DU POUVOIR DISCRIMINANT DES S-S DU LOGICIEL DE REPLIEMENT DES ARN EN SS, RNASUBOPT.	45
FIGURE 20. COURBE DU POUVOIR DISCRIMINANT DES S-S DU LOGICIEL DE REPLIEMENT DES ARN EN SS, MCFLASHFOLD.	45
FIGURE 21. COURBE DU POUVOIR DISCRIMINANT DE L'ENSEMBLE DE DONNÉES AYANT LA PLUS PETITE VALEUR POUR L'ÉQUATION 1, LE DISCRIMINANT.	46
FIGURE 22. VUE DÉTAILLÉE SUR LES S-S DANS RDV.....	52
FIGURE 23. VUE PRINCIPALE DE RNA DYNAMIC VIEWER (RDV).	53
FIGURE 24. PROTOCOLE D'APPRENTISSAGE GRAPHIQUE, PARTANT DES DONNÉES BRUTES JUSQU'À LEUR ANALYSE.	56

FIGURE 25. TABLE DE CONTINGENCE DE L'ÉTAT (PAIRÉ OU NON) DES NT. AVEC LEUR RÉACTIVITÉ CHIMIQUE.	57
FIGURE 26. COURBE DE ROC DU MODÈLE DE RNASS ET DU MODÈLE PAIRÉ / NON PAIRÉ.....	58
FIGURE 27. PRÉCISION EN FONCTION DU RAPPEL DU MODÈLE DE RNASS (EN BLEU) ET DE CELUI BASÉ SUR L'ÉTAT PAIRÉ NON PAIRÉ DES NT. (EN ROUGE).....	59
FIGURE 28. DISTRIBUTION DES NT. EN FONCTION DE LEUR ÉTAT (AXE VERTICAL), DE LEUR SCORE DE PRÉDICTION DE RNASS (AXE HORIZONTAL) ET DE LEUR RÉACTIVITÉ CHIMIQUE (COULEUR).	60

Liste des sigles

1m7	: 1-methyl-7-nitroisatoic anhydride
1m6	: 1-ethyl-6-nitroisatoic anhydride
ACP	: Amplification en chaîne par polymérase
API	: Application programming interface
ARN	: Acide ribonucléique
AUC	: Aire sous la courbe (<i>Area under the curve</i>)
NMIA	: N-methyl isatoic anhydride
Nt. :	: Nucléotide
MCN	: Motif cyclique nucléotidique
MFE	: Énergie libre minimum (<i>Minimum free energy</i>)
ROC	: Fonction d'efficacité du récepteur (<i>Receiver operating characteristic</i>)
RMDB	: RNA mapping data base
RMN	: Résonance magnétique nucléaire
SS	: Structure secondaire
S-S	: Sous-Structure
URL	: Localisateur uniforme de ressource (<i>Uniform Resource Locator</i>)

Liste des abréviations

3D	: Trois dimensions
A	: Adénosine
C	: Cytosine
U	: Uracile
G	: Guanine
DE	: Diversité d'ensemble (<i>ensemble diversity</i>)

À toi qui lis ce mémoire!

Remerciements

Mes remerciements vont à mes parents, mes frères, mon directeur de recherche, Dr François Major et aux quatre professeurs qui ont accepté d'être membre de mon jury : Gertraud Burger, Pascale Legault, Sylvie Hamel et Sergei Chteinberg. De plus, je tiens à remercier mes collègues de laboratoires, en particulier, Guillaume, Mathieu, Gabriel, Olivier, Zohra, Jordan, Roqaya, Blandine, Albert, Laurence, Thomas, Nathanael, Nicolas, Marc Frédéric, Paul, Maria, Julie R., Julie P., Mosen et le dernier, mais non le moindre Frank. De plus, je veux souligner l'aide que Patrick Gendron m'a apportée avec le serveur. Éline Meunier, la coordonnatrice des étudiants de bio-informatique, alias notre 2^e maman, m'a aussi beaucoup aidé. À vous tous, merci beaucoup!

Avant-propos

1.1 Avertissement

Les méthodes utilisées dans mon projet de recherches touchent plusieurs champs d'expertise. Pour faire l'analyse de la réactivité des nucléotides (nt.), j'ai dû créer des algorithmes robustes, bâtir des bases de données performantes et utiliser un serveur fourni par mon laboratoire de recherche.

Une solide base en informatique a donc été nécessaire. J'ai dû faire preuve de curiosité et de persévérance tout au long de l'écriture de ce mémoire. L'univers de la biologie moléculaire est très vaste et c'est pourquoi pendant toute ma maîtrise, j'ai tenté de garder les choses simples. J'ai voulu combiner plusieurs approches en un tout cohérent et facile d'utilisation.

Les méthodes à haut débit comme SHAPE-seq permettent de tester beaucoup d'hypothèses et les logiciels libres comme MCFlashfold et RNAsubopt peuvent être constamment amélioré et utilisé de plusieurs façons. Ces fonctionnalités en font des outils scientifiques vraiment précieux.

Aucune donnée n'a été torturée pendant le processus.

1.2 Mise en contexte

Ce travail a été effectué dans le laboratoire du Dr François Major à l'institut de recherche en immunologie et oncologie.

La bio-informatique moléculaire est une science récente et la détermination in silico de la structure des ARN est un vrai défi.

Actuellement, nous avons des outils informatiques pour représenter les ARN à plusieurs niveaux d'abstraction. Le niveau tout-atome est l'un des plus demandant en calcul, mais l'un des plus précis.

En 2018, une simulation par dynamique moléculaire de quelques microsecondes ne peut considérer que des structures de la grosseur d'une boucle de quelques nt., mais cela risque de changer rapidement.

Une autre limite de cette approche est que nul ne sait modéliser les interactions entre les ions magnésium et l'ARN. Il est bien connu que ces interactions sont essentielles au repliement de l'ARN [1].

La demande en calcul qu'entraîne cette simulation empêche les modélisations d'être faites en temps réel, un changement de paramètre entraîne un recalcul complet occupant un ordinateur pour plusieurs heures, voire plusieurs jours.

Grâce à mon logiciel, le visualisateur dynamique des ARN ou « RNA Dynamic Viewer » (RDV) on peut maintenant visualiser plusieurs structures secondaires (SS) d'un ARN (la SS de minimum d'énergie (MFE) et les SS sous-optimales) dans la même vue. Lorsqu'il est démarré sur un ordinateur équipé d'une carte graphique, la transition entre les différentes structures est fluide et cela permet d'avoir une idée de la distance physique entre deux ou plusieurs SS.

Avec ce logiciel nous pouvons aussi avoir une idée de la vraisemblance des SS en se basant sur la réactivité chimique relativement constante de certaines sous-structures (S-S).

Il offre un compromis entre la résolution, le temps de calcul et l'espace mémoire nécessaire. Son utilité principale est de permettre la création d'hypothèses menant à l'amélioration des prédictions des SS faites à partir d'une séquence sondée par le « 1-méthyl-7-nitroisatoic anhydride » (1m7).

Dans la littérature scientifique, l'évaluation d'une SS par le 1m7 est basée sur la flexibilité des nt., dérivée de leur état païré ou non païré [2, 3].

Avant 2011, la date de la sortie du protocole de SHAPE-seq [4], il était difficile d'obtenir des gros ensembles de données uniformes. Ces ensembles de données

permettent d'utiliser des algorithmes d'apprentissage machines. En 2014, comme le montre le graphique ci-dessous, le nombre de séquences sondées rendues publiques a connu une forte hausse, grâce au laboratoire du Dr Das et la *RNA Mapping Database* (RMDB) [5].

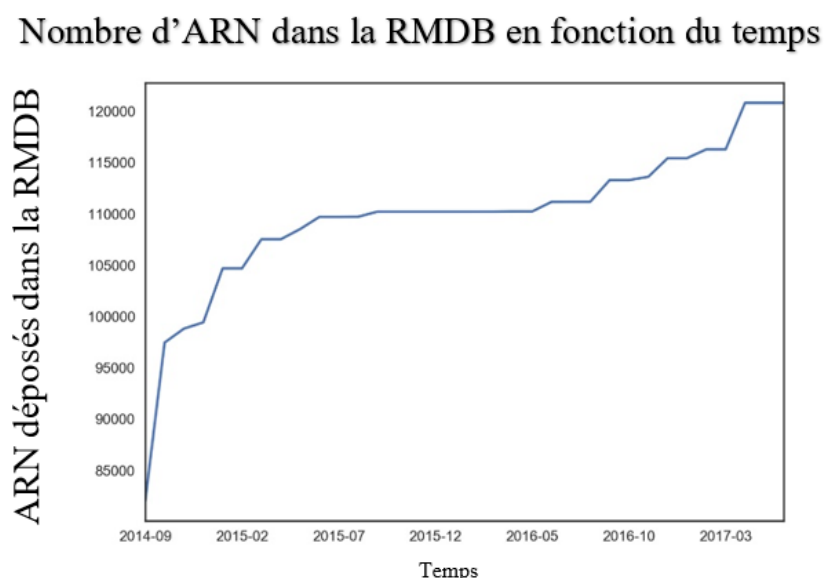


Figure 1. **Nombre de séquences d'ARN ajoutées à la RMDB entre septembre 2014 et mars 2017.** Figure modifiée de l'article: Updates to the RNA mapping database (RMDB), version 2. Nucleic Acids Research, 2017.

Mon approche est basée sur des S-S de la SS d'une grosseur de quelques nt.. Elle considère la majorité des prédictions, faite par les deux logiciels de détermination des SS utilisés, comme étant bonnes. Cette contrainte pose un défi de sélection des séquences abordé dans le chapitre 1. J'ai récolté les niveaux de réactivités chimiques d'un grand ensemble de données, c'est-à-dire, plusieurs millions de nt. provenant de plusieurs milliers d'ARN, je les ai compilées, pour ensuite prédire la réactivité des nt. provenant d'ARN non utilisés lors de l'apprentissage. Cela m'a permis d'évaluer mon

approche objectivement à l'aide des métriques couramment utilisées dans le domaine (courbe ROC, taux de vrais positifs, taux de faux positifs, précision, rappel, etc.)

En plus de donner un ordre d'exactitude des prédictions des SS, RNASS_v2 et son logiciel de visualisation, RDV permettent d'étudier les ARN dans le contexte de la réactivité chimique de façon générale. Ils sont disponibles librement sur *github*, dans le répertoire de Philippe Malric et dans celui du laboratoire du Dr Major.

1.3 Pourquoi étudier la structure de l'ARN

Le repliement de l'ARN est intrinsèquement lié à ses nombreuses fonctions. On peut penser aux ARN de transfert, dont la forme tridimensionnelle, en « L », est nécessaire pour qu'ils soient reconnus par le ribosome. Dans une cellule vivante, chaque ARN naît de la transcription d'un brin d'ADN. Ensuite, chaque molécule se replie pour adopter une ou plusieurs formes plus ou moins stables pouvant se transformer les unes en les autres sous certaines conditions. Ces molécules sont impliquées dans la majorité des processus de rétroaction cellulaires. Que ce soit au niveau de leur traduction en protéine ou directement. Les *riboswitchs* sont un bon exemple de régulation directe [6]. De façon indirecte, les micro-ARN (miARN) peuvent contrôler l'abondance de certaines protéines, en réprimant des ARN messagers (ARNm) [7]. L'effet éponge de certains ARN régule le niveau des miARN, le laboratoire du Dr Major est un pionnier dans le calcul des interactions des miARN in vivo [8].

De plus, un problème complexe comme celui abordé dans ce mémoire offre l'occasion d'innover en matière d'algorithme d'apprentissage machine. La possibilité d'analyser des données publiques permet à la communauté scientifique de se rassembler et d'échanger sur les meilleures pratiques, on peut penser en autres à « ETERNA » et sa convention annuelle : « Eternacon ».

Introduction

Comprendre la structure des ARN est essentiel à l'élucidation de son rôle dans la cellule. Les ARN agissent à plusieurs niveaux pour réguler l'expression génique. Ces molécules portent le message des gènes; elles peuvent réguler d'autres ARN par leurs représentants les plus petits, les micro-ARN et elles ont aussi la capacité de sentir leur environnement et ainsi moduler leurs fonctions de façon autonomes par des motifs nommés *riboswitch*.

Pour mieux saisir les subtilités des structures des ARN, une méthode a été développée : le sondage chimique des ARN. Parmi plusieurs méthodes de sondage chimique, j'ai choisi d'étudier la méthode nommée SHAPE. Au début des années 2010, des chercheurs ont combiné cette technique avec le séquençage haut débit. Le résultat fut nommé : SHAPE-seq, une méthode qui génère beaucoup de données.

Depuis 2014, le laboratoire du Dr Das à Stanford publie régulièrement des milliers d'ARN avec une valeur de réactivité et d'erreur pour presque tous les nt. des séquences sondées.

Dans mon projet de recherche, j'ai analysé ces données dans le but de pouvoir prédire la réactivité des nt. et pour mieux classer les structures secondaires (SS) produites par des logiciels de prédiction de SS.

Pour ce faire, il m'a fallu rassembler les connaissances sur les ARN sondés, visualiser et vérifier les prédictions. Dans le chapitre 3, j'ai comparé mon approche à celle basée sur l'état des nt. (paire ou non paire).

La décomposition des ARN en S-S ne permet pas de prédire parfaitement la réactivité des nt., mais cette technique est assez précise pour identifier des structures non conventionnelles, à la frontière de nos connaissances.

2.1 Historique

2.1.1 Début de l'étude des acides nucléiques

En 1869, Friedrich Miescher, découvre l'ADN [9]. 70 ans en plus tard, en 1939, Caspersson et Schultz utilisent un spectromètre pour établir qu'il y a des acides nucléiques dans le cytoplasme [10].

2.1.2 Détermination de la structure de l'ADN et de l'ARN

En 1953, Francis Crick et James Watson élucident, avec l'aide des données de Rosalind Franklin, un des mystères les plus grands du milieu du 20^e siècle, la structure de l'acide désoxyribonucléique (ADN) [11]. Les ARN font partie de la classe des acides nucléiques tout comme l'ADN et certaines de leurs structures sont semblables, les doubles hélices en sont de bons exemples.

En 1961, Monod et Jacob proposent que l'ARN soit un intermédiaire entre l'ADN et les protéines, ils auront raison [12]. Les mécanismes par lesquels les ARN jouent leurs rôles s'imposent alors comme questions de recherches.

En 1964, Robert W. Holley se sert d'une enzyme pour digérer un ARN de transfert et c'est ainsi qu'il découvre sa structure secondaire (SS) [13]. C'est le début du sondage enzymatique de l'ARN.

2.1.3 Prédiction de la structure secondaire des ARN

L'idée de Tinocco, en 1971, est de minimiser l'énergie de ces molécules pour trouver la structure la plus stable. Des règles simples sont appliquées à une structure pour évaluer son énergie. Elles comprennent : une énergie d'initiation des hélices, une énergie de propagation et une énergie pour les boucles [14].

En 1978, Nussinov propose un algorithme rapide pour déterminer la SS d'un ARN. Son nombre d'étapes grossit proportionnellement au cube de la longueur d'une

séquence de nt., autrement dit, il est dans l'ordre de complexité de $O(n^3)$, « n » étant le nombre de nt. de la séquence [15, 16]. Ensuite, il y a eu les algorithmes donnant les SS sous-optimales en 1984 et 1989 [17, 18].

En 1990, McCaskill imagine un algorithme pour obtenir la fonction de partition en $O(n^3)$ [19]. Cette fonction est utile pour connaître la probabilité de formation d'une paire de bases. En 1991, MC-SYM, un logiciel qui prédit la structure 3D d'une molécule est publié par le Dr François Major, mon directeur de recherche[20]. En 1994, ViennaRNA package voit le jour [21]. Par la suite, l'intérêt pour l'ARN n'a fait qu'augmenter.

En 1995, Turner publie un ensemble de données thermodynamiques sur les paires de bases et les boucles de l'ARN [22]. Ces données seront utiles pour des algorithmes tels que les proches voisins.

Une autre date marquante pour ce domaine est l'année 2008, MC-Fold une version moins performante de MCFlashfold, la version utilisée dans ce mémoire, est publié par le laboratoire du Dr Major [23].

2.2 Définition des termes et des concepts utilisés pour prédire la réactivité des nucléotides d'un ARN.

2.2.1 Les nucléotides

D'un point de vue chimique, l'ARN est un polymère de ribonucléotides. Il est composé dans la nature de quatre ribonucléotides communs et de plusieurs ribonucléotides modifiés. Les quatre bases communes de l'ARN sont l'adénine, la cytosine, l'uracile et la guanine. L'image en A de la figure 2 provient de [24] et l'image en B de la même figure provient de [25].

Un nucléotide (nt.)est toujours composé de trois parties, voir figure 2 :

1. Le groupement phosphate
2. Le ribose
3. La base azotée

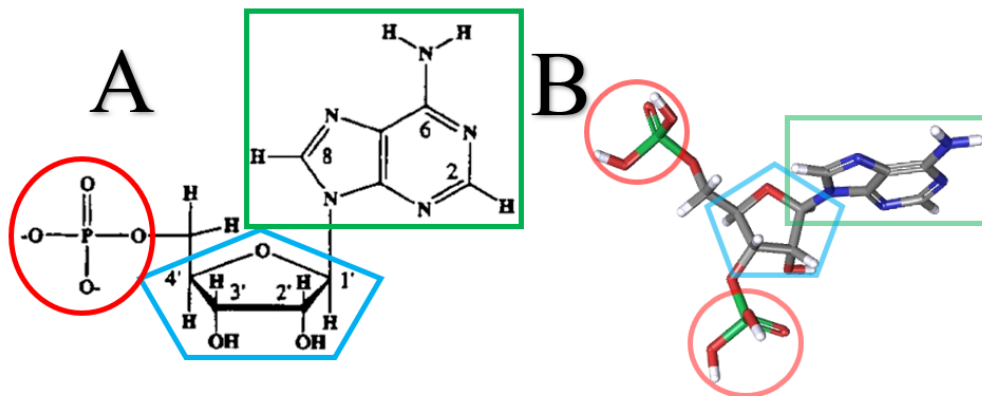


Figure 2. **Représentation d'un nucléotide, l'adénosine monophosphate (AMP).** A) Le cercle rouge, le pentagone bleu et le carré vert représentent le groupement phosphate, le ribose et la base azotée, respectivement. Cette image a été modifiée de l'article : Structural analysis of nucleic acid aptamers . Current opinion in chemical biology, 1997. 1: p. 32-46, écrit par Patel, D.J. B) Image 3D de A, avec en plus un groupe phosphate sur le carbone en 3' du ribose. Source : PubChem3D.

2.2.1.1 Le groupement phosphate

En ne comptant pas les atomes d'hydrogène, un groupement phosphate est composé de cinq atomes. Quatre atomes d'oxygène et un de phosphore. Chaque groupement phosphate donne une charge négative à l'ARN.

2.2.1.1 Le ribose

Le ribose est un pentose, un sucre composé de 20 atomes dans sa forme libre. Il a toujours cinq atomes de carbone. À l'état libre, il a cinq atomes d'oxygène et lorsqu'il est dans l'ARN, avec sa base et ses deux phosphates, il a deux atomes d'oxygène complètement à lui, deux oxygènes qu'il partage avec les deux groupes phosphate liés au carbone 3' et 5' et un atome d'oxygène qu'il partage avec sa base. Dans la figure ci-dessous, on observe un beta-D-ribose. Son image miroir, la forme « L » n'est pas présente dans la nature.

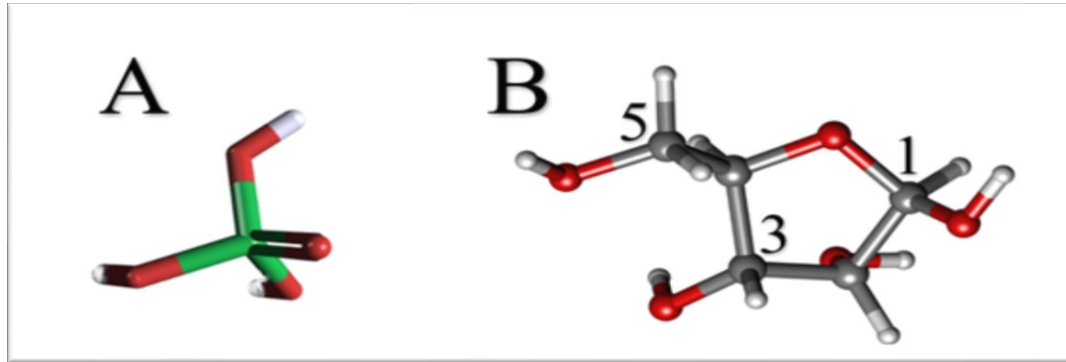
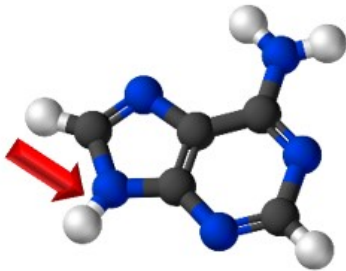


Figure 3. **Représentation en balles et bâtons (*balls and sticks*) d'un groupement phosphate et d'un beta-D-ribose.** Le phosphore est vert, les oxygènes sont rouges, les carbones gris et les hydrogènes blancs. **A)** L'atome de phosphore est au centre des atomes d'oxygène. Les hydrogènes en périphérie. Source : PubChem3D. **B)** Dans le polymère qu'est l'ARN, les groupes phosphates s'attachent aux carbones 3 et 5 du ribose. Le carbone 1, en beta, est lié à la base azotée. Cette image est disponible sur <https://commons.wikimedia.org/>, elle a été produite par, Marina Vladivostok, les positions (1, 3, 5) des atomes ont été ajoutés.

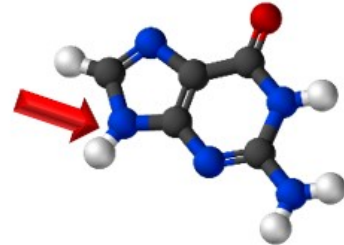
2.2.1.3 Les bases azotées

Les images suivantes viennent de Wikipédia, les atomes bleu, rouge, gris et blanc sont des azotes, des oxygènes, des carbones et des hydrogènes respectivement. La flèche rouge est le point d'attache du ribose. Comme mentionné plus haut, les ribonucléotides ont quatre représentants:

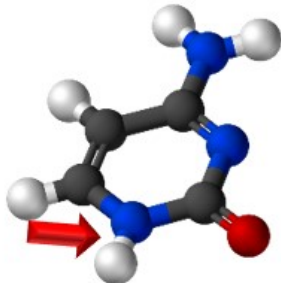
A) L'adénine



B) La guanine



C) La cytosine



D) L'uracile

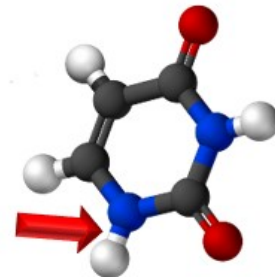


Figure 4. **Les quatre nucléotides des ARN.** A) L'adénine est une base azotée de la famille des purines. Son symbole est le A. B) La guanine est aussi une base azotée de la famille des purines. Son symbole est le G. C) La cytosine est une base azotée de la famille des pyrimidines. Son symbole est le C. D) Tout comme la cytosine, l'uracile est une base azotée de la famille des pyrimidines. Son symbole est le U.

2.2.1 La structure secondaire.

La structure secondaire (SS) d'un ARN est formée par l'ensemble des paires de bases d'un ARN. Ce concept est utile pour comprendre et expliquer, la forme des ARN. Un des composants majeurs des ARN est la double hélice.

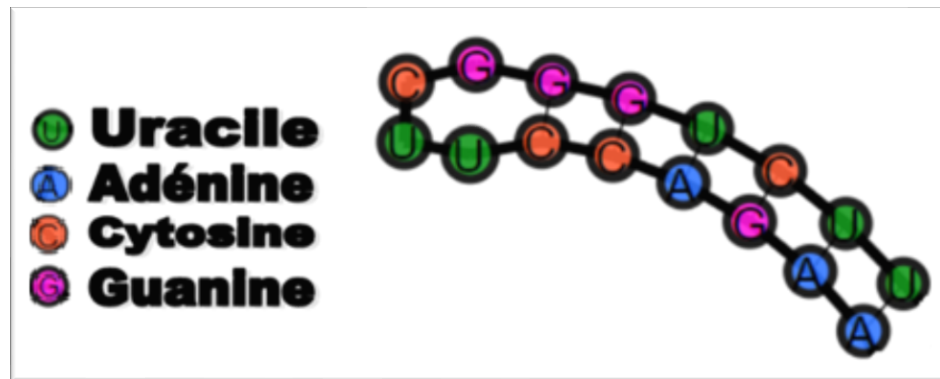


Figure 5. Exemple de visualisation d'une SS générée par le logiciel *RNA Dynamique Viewer*. (Voir chapitre 2)

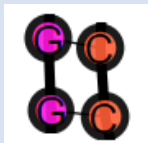
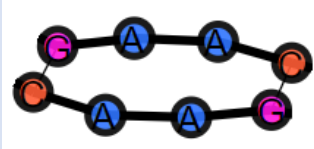
En général, on utilise un sous-ensemble des paires de bases répondant à deux critères, soit : l'absence de liens entre plus de deux nt. (sans compter les nt. adjacents liés de façon covalente) et l'absence de pseudo-nœuds. Les pseudo-nœuds sont des structures de la forme : $ux^1u y^1u x^2u y^2u$, où les « x » et « y » sont pairés ensemble : les 1 avec les 2, en exposant, de la même lettre et les « u » sont des nt. non pairés ou rien du tout [26]). Avec ces contraintes, le problème de la détermination de la SS peut être résolu avec la programmation dynamique en $O(n^3)$, sans ces contraintes le problème peut croître de façon exponentielle en fonction de la longueur de l'ARN. Il faut garder en tête que la majorité des liens entre les nt. sont compatibles avec ces critères, mais que les SS ainsi créés ne peuvent pas expliquer l'entière des phénomènes de réactivité chimique. La figure 5 est un exemple de tige suivi d'une boucle. De façon générale, les SS sont définies par un réseau composé de paires de bases et de liens covalents entre les nt.. Les liens entre les bases ont une énergie libre de Gibbs plus élevée que les liens

covalents. Parmi les paires de base, c'est la paire : « cytosine – guanine » qui a l'énergie libre de Gibbs la plus basse suivi de la paire : « adénine – uracile »[27].

2.2.2 Les cycles

Un cycle est une S-S de la SS des ARN. Un ARN, est formé de boucles, de doubles hélices, de jonctions et de motifs tertiaires. Dans la suite de ce mémoire, seul les deux premières S-S sont étudiées La définition des cycles s'appuie sur les le concept de paires de bases. Par exemple, on désigne par « 2_2 », le cycle le plus fréquent. Il est fait de deux paires de bases adjacentes et il forme un cycle de quatre nt.. Le tableau I présente deux cycles, le « 2_2 » est le composant principal des doubles hélices et le « 4_4 » est un renflement interne.

Tableau I. Représentation de deux cycles simples

Cycle	2_2	4_4
Figure		
Identifiant	2_2_GG_CC	4_4_GAAC-GAAC

Un cycle « 3_2 » est formé d'un nucléotide non pairé au centre de deux nt. pairés avec les deux nt. adjacents sur l'autre brin de l'ARN. Communément, on appelle cette structure, un renflement d'un nucléotide (un *buldge*).

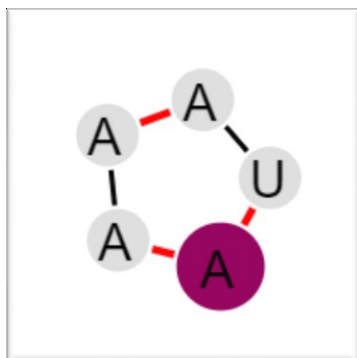


Figure 6. **Un cycle « 3_2 ».** Les arêtes rouges sont des liens covalents entre les nt. et les liens noirs sont les liens formant les paires de bases. Dans cet exemple, le nucléotide en rouge est le nucléotide d'intérêt, sa couleur nous renseigne sur sa réactivité moyenne (voir chapitre 2).

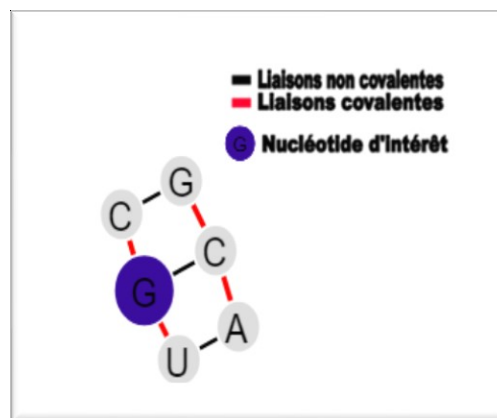
2.2.3 Les sous-structures

Pour caractériser un nt. deux sortes de sous-structures (S-S) ont été utilisées :

1. Des cycles simples. C'est-à-dire, lorsque le nucléotide est pairé, deux entrées sont ajoutées dans la base de données (une pour le cycle en 5' et l'autre pour le cycle en 3'). Ces motifs sont plus petits. Dans l'exemple de la figure 7, la guanine fait partie de deux cycles.
2. Une S-S englobant le nucléotide. C'est-à-dire qu'il y a invariablement une entrée peu importe l'état pairé ou non du nucléotide.

Le chevauchement de deux cycles est illustré dans la figure 7 et 8.

Figure 7. **Une S-S formée de deux cycles accolés.** L'identifiant complet de cette S-S est : « 2_2-UG-CA_pos_1&2_2-GC-GC_pos_0 ». Le nt. d'intérêt est le deuxième nucléotide du cycle « UG-CA » ou le premier nt. du cycle : « GC-GC », l'identifiant permet de comparer les nt. se trouvant dans la même S-S sur d'autres ARN.



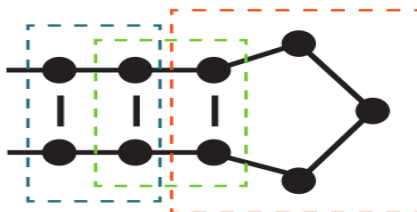


Figure 8. **Représentation de trois cycles se chevauchant.** Dans cette figure, le chevauchement des cycles est évident. La boîte bleue et la boîte verte entourent un cycle « 2_2 ». La boîte rouge entoure un cycle à cinq nt., un « 5 ». Lorsqu'on s'intéresse à un nucléotide en particulier, les nt. pairés se trouvent à la jonction de deux cycles tandis que les nt. non pairés font partie d'un seul cycle. Cette figure est tirée du manuel d'instruction de MCFlashfold (2015, Dallaire, P. and F. Major, *MCFlashfold (mcff version 34) user manual*). Les boîtes ont été ajoutées.

Dans ce mémoire, une chaîne de caractère identifie les S-S. Cet identifiant est un mini langage en soi au sens informatique du terme. Le diagramme à rail de l'expression régulière qui lui correspond a été placé dans la figure suivante.

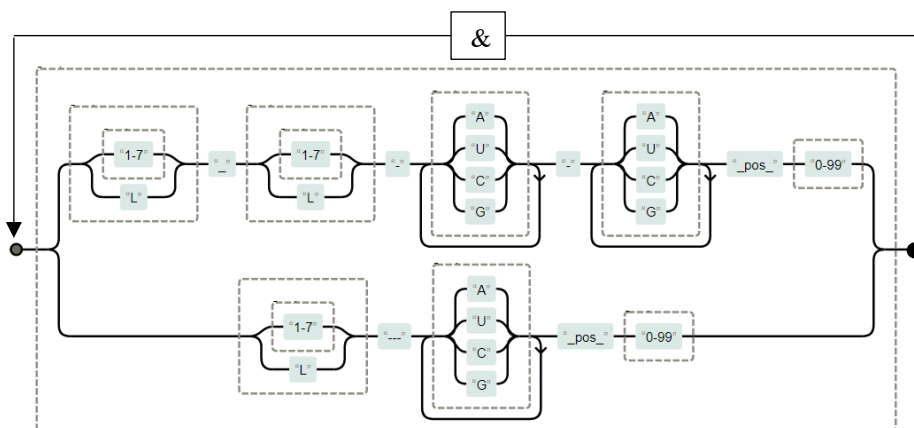


Figure 9. **Diagramme à rail de l'expression régulière associée à l'identifiant des S-S.** Un « & » est placé entre les deux chaînes de caractères lorsque le nucléotide est au centre de deux cycles.

Les deux sortes de S-S utilisent le même identifiant de base (les cycles simples et les cycles composés).

Cette page et la prochaine décrivent textuellement l'identifiant.

2.2.4 L'identifiant unique d'un nucléotide à l'intérieur d'une S-S

L'identifiant d'un nucléotide à l'intérieur d'une S-S est composé d'un ou deux « cycle(s) augmenté(s) », lorsqu'il y en a deux, un « & » les séparent. Chaque « cycle augmenté » est formé d'un « type », d'une ou deux séquences et de la position du nucléotide d'intérêt. Le terme augmenté réfère à l'addition de la position du nucléotide dans le cycle, cette précision est utile pour caractériser la réactivité d'un nucléotide dans une S-S.

2.2.4.1 Le type

Dans tous les cas, la chaîne de caractère débute toujours par le type du cycle (« 2_2 », « 3_2 », « 4_4 », etc.).

Lorsque le cycle a deux brins, il est délimité par deux paires de bases (deux liaisons non covalentes), le type est composé de deux chiffres représentant le nombre de nt. entre les deux paires de bases plus les nt. des paires de bases. Il débute par le brin du nucléotide d'intérêt.

Les boucles sont traitées de la même manière, mais elles ont un seul brin. Précisément, pour les boucles, le type du cycle a seulement un chiffre. Ce chiffre est le nombre de nt. libres de la boucle plus les deux derniers nt. de l'hélice.

Seuls les cycles ayant moins de sept nt. sur chaque brin ont un chiffre, les autres ont un « L ». Deux raisons expliquent ce choix d'implémentation.

1. Il n'y a pas beaucoup d'occurrences d'une grosse S-S.
2. L'espace de la base de données est précieuse, ses performances dépendent du nombre d'entrées différentes.

2.2.4.2 La séquence

Pour les cycles à deux brins, le type est suivi par les séquences, en commençant, aussi, par le brin d'où provient le nt. d'intérêt. Les séquences sont de la longueur des chiffres du type ou si elles sont égales ou plus grandes à sept, elles sont remplacées par la lettre « L ».

Pour les boucles, la séquence débute par le nucléotide pairé en 5' et se termine par le nucléotide pairé en 3'. Une fois encore, lorsque le nombre de nt. est égal ou supérieur à sept, la lettre « L » remplace la séquence.

2.2.4.3 La position du nucléotide

Finalement, on ajoute à l'identifiant la position du nucléotide. Pour les cycles composés de deux brins, le nt. se trouve sur le premier brin. Ici il faut noter qu'on ne prend pas en compte le côté du brin sur lequel se trouve le nt. (5' ou 3') lors de leur extraction de l'ARN. Pour les boucles, il n'y a pas d'ambiguïté à ce niveau.

Dans les deux cas (les S-S à deux brins ou boucles), le premier nucléotide en 5' a comme numéro de position le 0 et les autres suivent. Le dernier nt. a donc la position « n » - 1, « n » étant le nombre de nt. du brin.

2.2.4.4 Appartenance d'un nucléotide à un cycle

Les nt. pairés font partie de deux cycles, tandis que les nt. non pairés font partie d'un seul cycle par SS. Les nt. de début ou de fin de séquences pour lesquels aucun voisin n'est pairé d'un des deux côtés ne font pas partie d'un cycle à proprement parler, mais il est utile d'identifier ces caractéristiques pour prédire la réactivité des nt.. L'algorithme qui permet d'extraire les cycles d'une SS est expliqué dans le chapitre 1.

2.2.5 La prédiction des structures secondaires

Pour comprendre et prédire la réactivité des nt. à partir d'une séquence de nucléotide, il faut avoir une idée de sa structure. Plusieurs méthodes ont été

développées pour prédire la ou les structures secondaires d'un ARN à partir de sa/leur séquence. Dans ce mémoire, j'utilise deux d'entre elles : la méthode des proches voisins et celle de l'échantillonnage des structures connues.

2.2.5.1 Les proches voisins (RNAsubopt)

Une des deux méthodes de repliement des ARN en SS utilisé dans ce projet de recherche est la méthode des proches voisins.

Cette méthode est nommée « *nearest neighbour* » en anglais et a été développée par Nussinov et coll. en 1978 [15]. Zucker et Stiegler l'ont implémentée et placée dans un programme nommé mfold par la suite [28].

Dans cet algorithme, les duos de paires de bases adjacentes, les boucles internes, les boucles de fin d'hélice et les renflements obtiennent une énergie qui contribue à l'énergie totale de la molécule. L'idée, tout comme dans beaucoup d'autres algorithmes, est de former la SS ayant la plus petite énergie possible. Cette méthode a été optimisée et est très utilisée pour déterminer la température de fusion (ou de fonte) de deux acides nucléiques (ADN ou ARN) lors d'une 'amplification en chaîne par polymérase (ACP). Un logiciel produisant un ensemble de SS en utilisant cet algorithme a été intégré à mon approche, il se nomme : *RNAsubopt* [30].

2.2.5.2 Échantillonnage à partir des structures 3D connues (MCFlashfold)

La deuxième méthode de repliement des ARN utilisés est une méthode développée dans le laboratoire du Dr Major.

Elle consiste à échantillonner les motifs cycliques nucléotidiques (MCN) d'un ensemble d'ARN dont les structures 3D sont connues. L'ARN est transformé en un réseau, où les nœuds sont des nt. et les liens sont soit des paires de bases ou des liens covalents médiés par un groupement phosphate entre les sucres des nt..

La fréquence d'observation d'un motif donné est convertie en probabilité et ces probabilités sont par la suite utilisées pour prédire les SS. L'algorithme est offert dans le matériel supplémentaire de l'article MC-Fold / MC-SYM pipeline [23].

MCFlashfold, la version utilisée, est une amélioration de MC-fold, créé par un ancien étudiant du Dr Major, Dr Paul Dallaire. MC-fold a été créé par Dr Marc Parisien.

2.2.6 Le sondage chimique de l'ARN

Une des méthodes fructueuses dans l'étude des structures des ARN est le sondage chimique des ARN. Les protocoles de sondages chimiques existent depuis longtemps. Ils ont été optimisés et ils produisent des résultats abondants grâce aux technologies de séquençages de nouvelle génération.

Le principe est le suivant. L'ARN est synthétisé en laboratoire. Il est placé en contact avec un agent chimique qui le modifie de façon spécifique. La réactivité d'un nt. est liée à la flexibilité de l'ARN près de lui et à son accessibilité à l'agent chimique [29].

RNAstructure, un logiciel de détermination de la SS des ARN, incorpore les données de réactivité chimiques en appliquant des contraintes sur la SS. Selon cette méthode, si le nucléotide réagit, il ne peut être païré à moins d'être dans une paire « guanine - uracile » (G-U) ou de suivre une paire G-U [30]. Pour Turner, Zucker et Mathews, en plus des règles déjà décrites, les nt. peuvent être dans une paire « adénine - uracile » (A-U) et « guanine - cytosine » (G-C) à la fin d'une hélice [31]. D'autres règles existent : Stadler ajoute des bonus d'énergie aux nt. réactifs, au lieu de déterminer complètement leur état (païré ou non). Ces contraintes sont appelées des contraintes souples, par opposition aux contraintes dures ou rigides citées précédemment [32].

Les nt. peu réactifs sont inaccessibles à l'agent modificateur. Cette inaccessibilité est due soit au païrage du nucléotide soit à la structure tertiaire de l'ARN, soit à une conformation locale précise.

Le réactif et les conditions des expériences sont déterminants dans les résultats finaux. Dans ce projet de recherche, seul le réactif nommé « 1m7 » a été étudié. Toutes les données proviennent de la RMDB et sont associées à l'ensemble de données « *ETERNA* ».

Le nombre de modifications par ARN est contrôlé par la concentration du réactif et par son temps de contact avec l'ARN. Les protocoles les plus récents prennent en compte le décrochage de l'ARN en soustrayant les valeurs obtenues par un contrôle négatif. Dans le but de prendre en compte l'erreur due à un nucléotide très réactif, plusieurs sondages à différentes concentrations du modificateur sont nécessaires, ce phénomène se nomme en anglais : « *over modification* » [33]. Ensuite une enzyme, la rétrotranscriptase transcrit l'ARN en ADN. Lors de cette étape, la rétrotranscriptase s'arrête aux bases modifiées. À l'aide des technologies de séquençages, on obtient un décompte de toutes les sous-séquences de la séquence. Le logiciel « Mapseeker » établit ensuite une correspondance entre le nombre des molécules d'une certaine séquence et le lieu où l'ajout s'est produit [34]. Beaucoup de détails sur la standardisation de ces expériences sont disponibles dans le matériel supplémentaire de l'article [33].

2.2.6.1 SHAPE-seq

Le laboratoire du Dr Weeks a développé une méthode nommée SHAPE qui est l'acronyme de « *Selective 2' Hydroxyl Acylation analyzed by Primer Extension* ». Cette méthode est composée de trois réactifs : le « 1-méthyl-7-nitroisatoic anhydride » (1m7), le « 1-éthyl-6-nitroisatoic anhydride » (1m6) et le « N-méthyl isatoic anhydride » (NMIA). Par la suite, le laboratoire du Dr Lucks a développé une autre méthode basée sur SHAPE qui a permis d'augmenter le nombre d'ARN sondés [35]. Le Dr Das et ses collègues ont perfectionné cette méthode et ils ont créé une base de données pour obtenir les ARN et leurs valeurs de réactivité. Cette base de données se

nomme : la RMDB, pour « *RNA mapping database* ». Comme il a été mentionné plus haut, les ARN analysés dans ce mémoire ont été sondés avec le 1m7.

2.2.6.2 Le 1-methyl-7-nitroisatoic anhydride (1m7)

La figure 10 montre les réactifs et les produits de la réaction chimique entre le 1m7 et une adénine. Une molécule de dioxyde de carbone est créée lors de la réaction. C'est le groupement hydroxyle du carbone 2 du ribose qui réagit avec le 1m7. La sélectivité et le mécanisme de cette réaction ne sont pas complètement compris [36]. Récemment, une équipe a montré à l'aide de la dynamique moléculaire que la conformation locale du nt. et spécialement de son sucre est importante dans ce mécanisme [37].

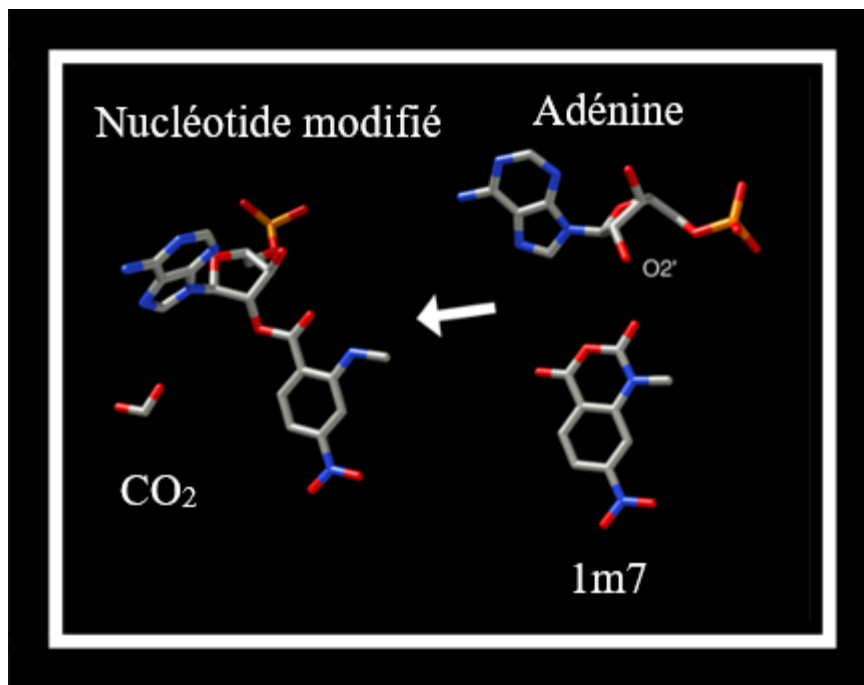


Figure 10. **Réaction chimique du 1M7 avec le 2'OH du sucre d'un nucléotide.** Cette image provient du wiki d'Eterna, un jeu vidéo consacré à la structure des ARN : <http://eternawiki.org/wiki/index.php5/1M7>. Elle a été confectionnée par Omei.

Chapitre 1: RNASS_v2 : Obtenir la réactivité des S-S

RNASS_v2 est un script écrit en dans le langage informatique *python*. Il est spécialement conçu, pour faire le lien entre les données de sondages chimiques de l'ARN et la SS produite par les logiciels *MCFold* et *RNAsubopt*. Ce script prend en entrée une liste d'ARN. Il produit en sortie un fichier JSON. Ces fichiers sont ensuite utilisés pour peupler une base de données qui sert à prédire la réactivité des nt.. Dans ce chapitre, je décris les étapes de son fonctionnement. Il a deux modes :

1. Le mode : apprentissage
2. Le mode : prédiction

3.1 De l'obtention des données à la prédiction discrète

3.1.1 Extraction des données

La procédure pour obtenir expérimentalement les données de la RMDB et les traitements numériques subséquents sont bien expliqués dans la documentation [33, 34] et sur le forum d'« *ETERNA* » [38]. Pour les données d'« *ETERNA* », une interface de programmation (API) a été mise en place pour que la communauté puisse les obtenir facilement. Il suffit de trouver la liste des « laboratoires » et de faire une requête à l'API (voir l'annexe). Les « laboratoires » sont des ensembles de données créés dans un but précis, ce but peut être l'étude d'une boucle, d'un renflement ou de toutes autres caractéristiques. Plusieurs informations sont disponibles sur les « laboratoires », nous retrouvons entre autres : leurs auteurs, leur identifiant numérique, la date de leur création, leur titre et une description. Par exemple, le tableau de la page suivante correspond à l'expérience : « *ETERNA_R00_0000* »

Tableau II. Exemple d'information sur les expériences de la RMDB.
(ETERNA_R00_0000)

Nom	Valeur
Authors	Ann Kladwang, EteRNA players, Rhiju Das.
Comments	Output of MAPseeker v1.2 from data: 122112_RD_EteRNAPlayerProjects_1M7test
Construct_count	2059
Creation_date	06/10/15
Modifier	1M7
Chemical	MgCl2:10mM,HEPES:50mM(pH8.0)
Data_count	162582
Description	The first 'cloud lab' experiments -- EteRNA player designs from 2012, testing a wide range of hypotheses.
Name	EteRNA Cloud Lab
Rmdb_id	ETERNA_R00_0000

Un script *python* nommé : eternaTSV.py utilise cet API et produit une liste d'ARN avec leur séquence, les données de réactivité chimique et d'autres informations reliées à l'expérience qui les a créées.

Il est aussi possible de télécharger les fichiers *RDAT* de la base de données, RMDB. Ces fichiers viennent avec un script *python* qui se nomme *handler.py*. Chose surprenante le vecteur de réactivité n'est pas de la même longueur que la séquence elle-même. On doit phaser le vecteur de réactivité avec un attribut nommé : « *offset* ». Ceci s'explique par le fait que l'ARN possède une séquence qui permet son identification lors du séquençage. Cette séquence n'a pas de données de réactivité.

Carte de chaleur (*heatmap*)

L'image ci-dessous représente les 2059 séquences sondées de l'expérience : « ETERNA_R00_0000 ». Chaque séquence vient avec une structure dont la réactivité chimique est connue pour atténuer les variations entre les séquences. Les ARN sont aussi munis d'un identifiant sous forme de chaîne nucléotidique tel que décrit dans le protocole du SHAPE-seq [35].

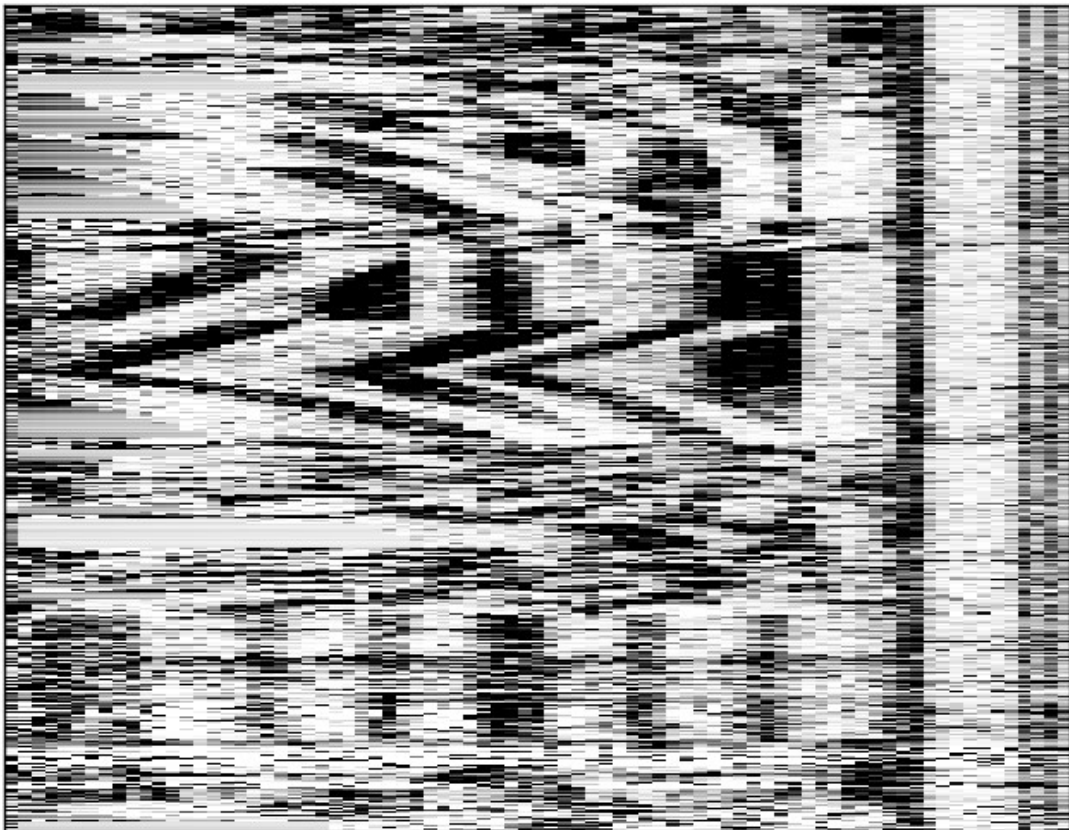


Figure 11. **Carte de chaleur de l'expérience : « ETERNA_R00_0000 ».** Chaque ligne représente une séquence et les colonnes représentent la réactivité associée à chaque nucléotide. Cette figure provient du site web de la RMDB : https://rmdb.stanford.edu/detail/ETERNA_R00_0000.

Les figures suivantes démontrent que l'algorithme qui aligne la séquence sur le vecteur de réactivité est cohérent avec les données de la RMDB. Chacune des trois figures a deux SS, celles de gauche proviennent de RDV et celles de droite proviennent du site web de la RMDB. On voit que les données de réactivités correspondent.

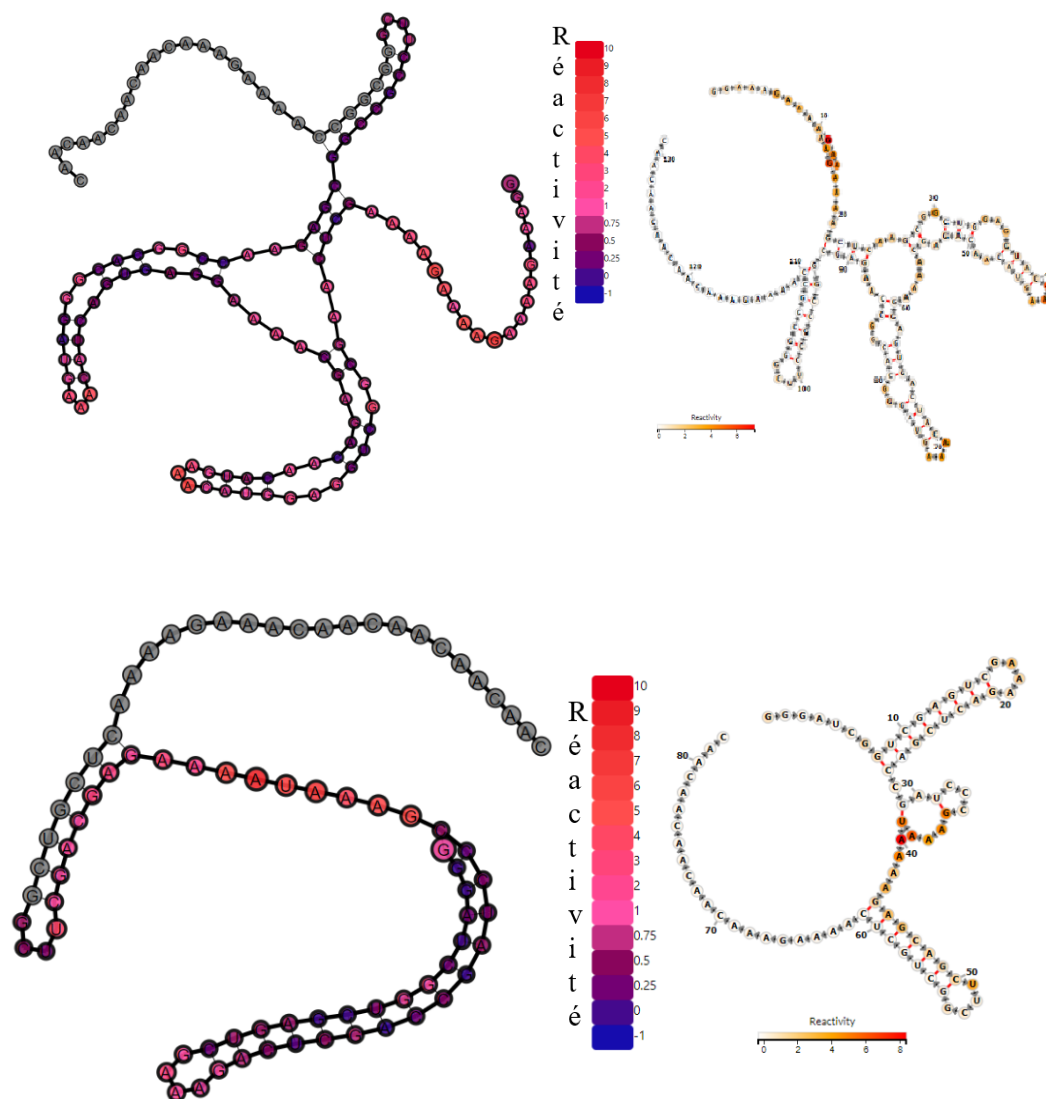


Figure 12. **Les données de réactivités sont également phasées pour RDV et le site web de la RMDB.** Bien que les structures secondaires soient différentes, les données de réactivité chimique sont identiques (voir la figure suivante).

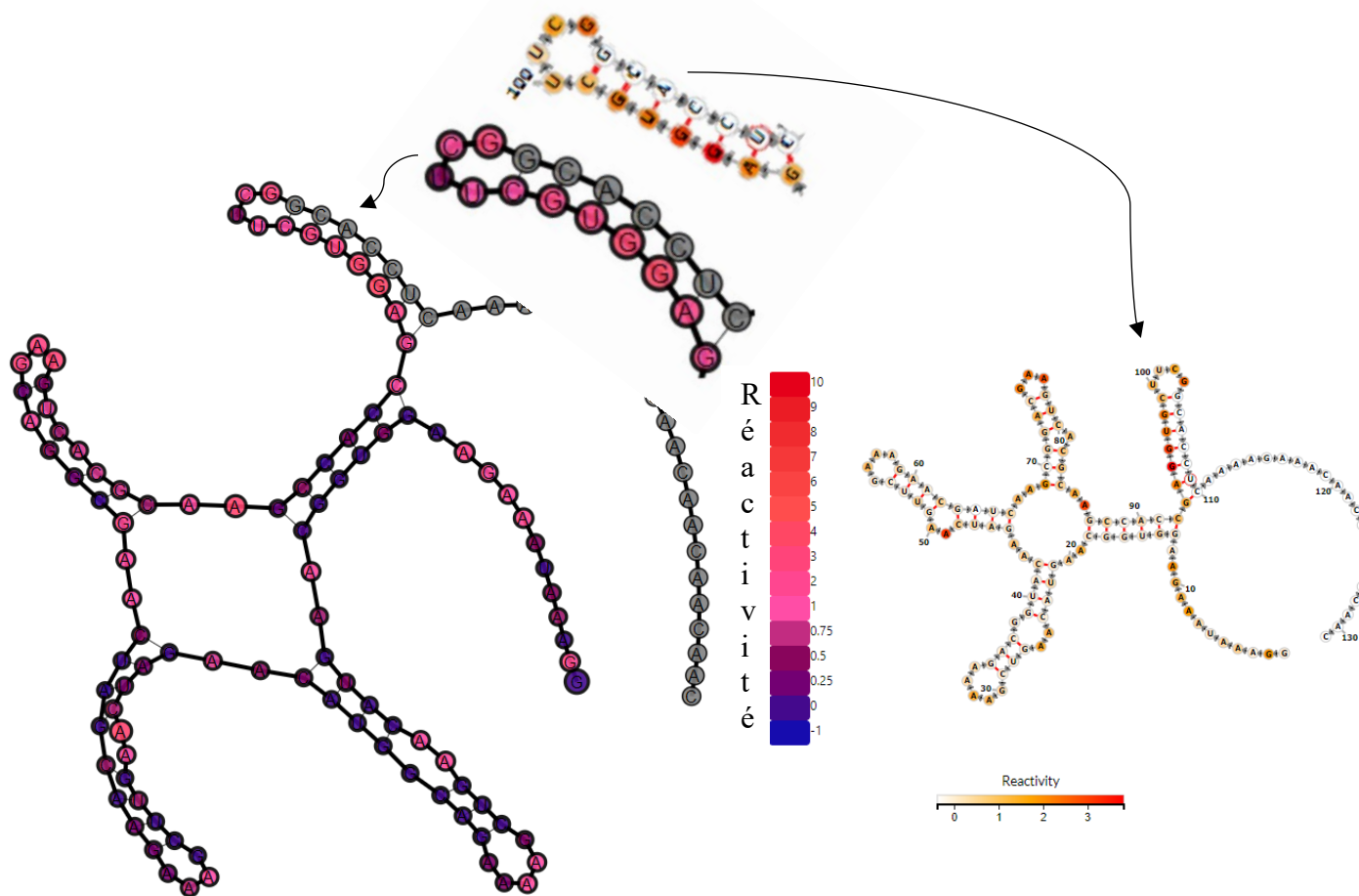


Figure 13. Comparaison entre une structure produite par *RNA Dynamic Viewer* et son homologue pris de la page web de la *RMDB*. Les deux échelles de réactivité permettent d'avoir une idée de la réactivité des nt.. Un nucléotide gris ou blanc n'a pas de valeur de réactivité. Les deux structures du haut (plus grosses) correspondent à la dernière double hélice de la SS. La guanine est en 5' et la cytosine en 3'.

Algorithme d'identification des cycles

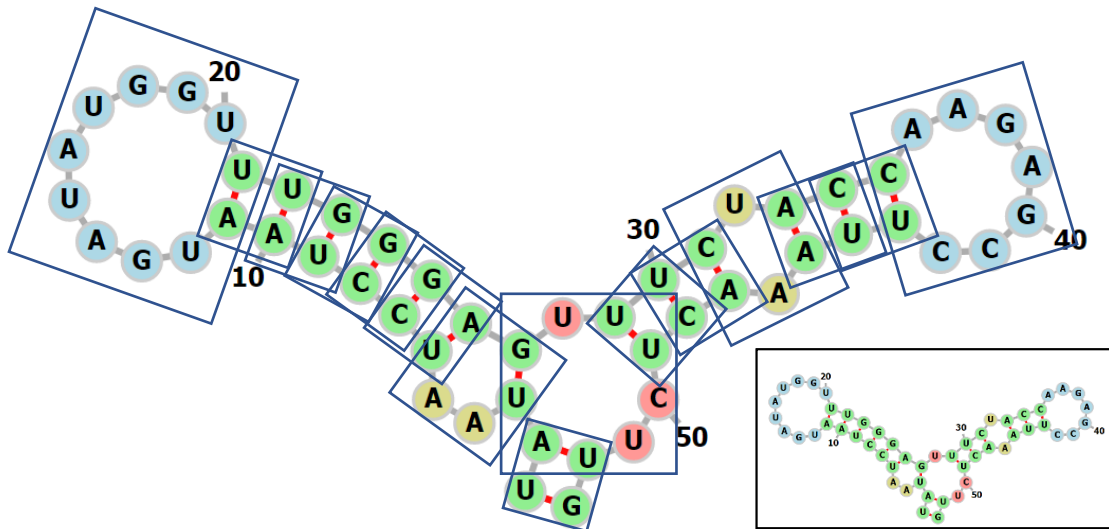


Figure 14. **SS provenant de FORNA.** Les doubles hélices sont vertes, les renflements sont rouges et les boucles terminales et intérieures bleus et jaunes respectivement. Les cycles sont placés dans des boîtes. La figure originale est en bas à droite.

La première étape consiste à rechercher les voisins pairés du nucléotide, on les nommera : le voisin gauche (en 5') et le voisin droit (en 3'). Ensuite, l'algorithme se divise en deux branches. Soit le nucléotide est pairé ou il est non pairé. Lorsqu'il est pairé, il fera partie d'un cycle comportant deux brins, la longueur de chaque brin est déterminée et ajoutée au début de la chaîne de caractères de sortie.

Lorsque le nucléotide est non pairé, deux autres branches sont accessibles. Soit, il se trouve dans une boucle, dans ce cas, le partenaire du premier voisin en 5' pairé et le premier voisin en 3' pairé sont le même nt, soit que le nt est dans une boucle interne ou un renflement, les deux derniers cas sont gérés de la même façon.

Dans tous les cas, lorsqu'il est non pairé, il appartient à un seul cycle, la figure 14 nous le démontre clairement.

Dans le cas de la boucle, on comptabilise le nombre de nt. non pairés, plus les deux premiers et dans les deux autres cas, c'est la distance en nt. entre les deux voisins pairés et leurs partenaires qui est comptabilisé.

Finalement, on identifie la S-S par la séquence des ou du brin(s) le composant et on ajoute la position du nucléotide. La valeur de la position débute toujours par 0.

Ceci est fait pour chaque SS généré par les deux logiciels. Cet algorithme fonctionne avec n'importe quel logiciel fournissant une ou des SS sous forme de chaîne de caractères ou sous forme de liste de paires de bases. La figure suivante illustre cet algorithme.

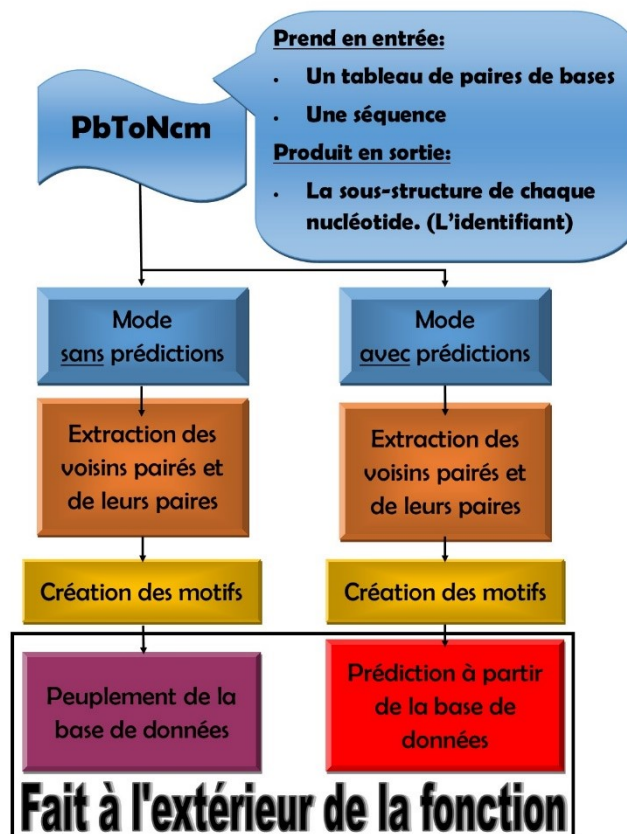


Figure 15. **Illustration de l'algorithme d'identification des S-S.** Pour faire ses prédictions, cette fonction a besoin d'une base de données donnant la réactivité des S-S (un tuple : identifiant / valeur).

3.2 Création des ensembles de données

Après la première analyse de mon algorithme, les ARN dont le score de prédiction moyen bas était systématiquement des ARN ayant un contenu en adénine ou un score de réactivité moyen élevé. Pour vérifier cette hypothèse, plusieurs ensembles de données ont été générés.

Différents seuils ont été testés. Ces paramètres sont : le score de réactivité moyen, une valeur nommée « *signal to noise* », fournie avec chaque ARN sondé et la diversité d'ensemble des SS des ARN. Le paramètre : proportion en adénine a été fixé à au plus 50%. Tout cela dans le but de réduire le bruit apporté par l'incertitude au niveau de l'expérience de sondage chimique et de la prédiction des SS.

Pour comparer la valeur prédictive anticipée des différents ensembles de données créés à partir des seuils sur ces paramètres, une formule ayant comme caractéristique de donner un résultat bas lorsque l'ensemble sépare bien les S-S a été utilisée. Chaque S-S de chaque ARN est pris en compte.

Dit autrement, les paramètres ont été optimisés sur leur pouvoir à séparer les S-S en deux groupes, les S-S réactives et celles qui ne sont pas. La formule est expliquée plus loin.

Les prochaines figures donnent la distribution de :

- la diversité d'ensemble (figure 16)
- la réactivité moyenne (figure 17)
- le « *signal to noise* » (figure 18)

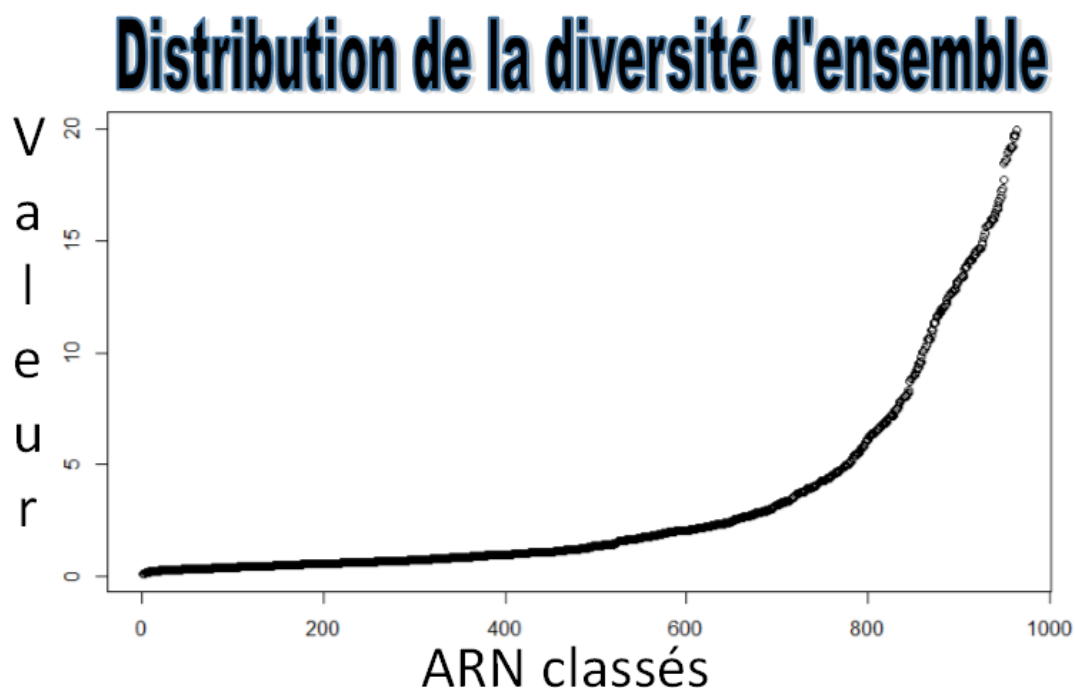


Figure 16. **La distribution de la diversité d'ensemble de 1000 ARN provenant de la RMDB.** Sur l'axe vertical, les valeurs de la diversité d'ensemble se situent entre 0 et 20. Une valeur basse signifie que toutes les structures sous-optimales ressemblent à la SS optimales, la MFE. Sur l'axe horizontal, 1000 ARN ont été classés en ordre croissant.

Puisque seule la SS d'énergie minimale a été étudiée, il est important qu'elle reflète bien la structure de l'ARN. Pour augmenter les chances que la structure de minimum d'énergie soit la bonne, un seuil a été placé sur la diversité d'ensemble (DE). Tous les ARN ayant une DE inférieure au paramètre testé à chaque itération ont été repliés et analysés.

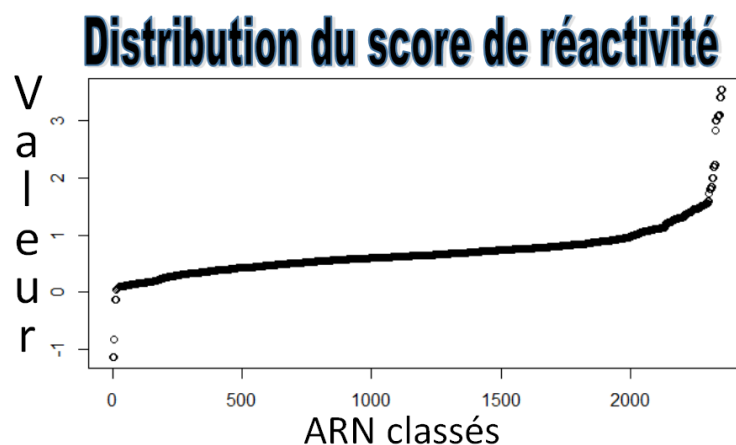


Figure 17. **La distribution de la réactivité moyenne de 2500 ARN provenant de la RMDB.** Sur l'axe vertical, les valeurs du score de réactivité moyen situent entre -1.5 et 4. Ces valeurs sont normalisées sur une structure connue ajoutée à la séquence. Sur l'axe horizontal, près de 2500 ARN ont été classés en ordre croissant de réactivité moyenne.

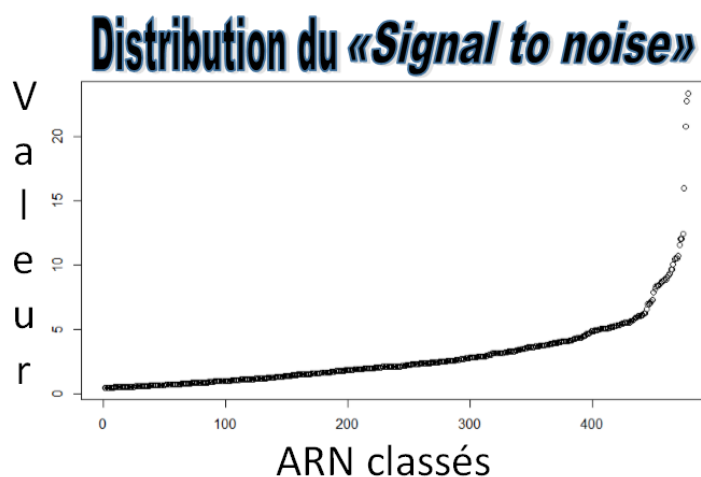


Figure 18. **La distribution du « signal to noise » de 2500 ARN provenant de la RMDB.** Sur l'axe vertical, les valeurs du score de réactivité moyen se situent entre 0 et 25. Une valeur au-dessus de 1 est considérée comme bonne. Plus la valeur est basse moins le signal est fort en comparaison à l'erreur. Sur l'axe horizontal, près de 500 ARN provenant de la RMDB ont été classés en ordre croissant.

L'équation à optimiser

La prochaine section présente une équation utilisée pour mesurer le pouvoir d'un ensemble de paramètres à classer la réactivité des nt..

L'ensemble de paramètres est utilisé pour générer un ensemble d'ARN. Les nt. de ces ARN sont classés en trois catégories :

Tableau III. Catégories de nt. en fonction de leur réactivité

Les nt. de basses réactivités	Réactivité $< 0,5$	Low
Les nt. de moyennes réactivités	$0.5 < \text{Réactivité} < 1.5$	Bg
Les nt. de hautes réactivités	$1.5 < \text{Réactivité}$	Hi

Explication de l'équation

Les S-S sont compilées dans une base de données en prenant en compte la position du nucléotide d'intérêt dans la S-S et sa réactivité classée dans l'une des catégories du tableau ci-dessus. L'occurrence des S-S est comptée par catégorie. Les nt. ayant une réactivité élevée sont désignés par : $o_{Hi}(s-s)$, tandis que ceux ayant une réactivité basse sont désignés par : $o_{Low}(s-s)$.

La différence entre l'occurrence des S-S élevées ($o_{Hi}(s-s)$) et basses ($o_{Low}(s-s)$) est calculée et elle est divisée par la somme des deux occurrences. À partir d'ici les nt. d'une S-S ayant que des valeurs élevées obtiennent une valeur de 1, tandis que les nt. d'une S-S ayant que des valeurs basses obtiennent une valeur de -1. Toutes les valeurs intermédiaires sont possibles. Dans les figures représentant le pouvoir discriminant théorique des ensembles de données (figure 19, 20 et 21), les points représentent cette valeur.

La valeur absolue (abs) de ces valeurs est prise et on la soustrait à 1 pour qu'une S-S ayant constamment des valeurs de réactivité de la même catégorie ai un score de 0. Pour favoriser les basses valeurs de façon non linéaire, le carré de cette valeur est pris. Une démonstration de ce que donne cette dernière manipulation est faite à l'aide du tableau IV (voir l'explication en dessous).

Tableau IV. Propriété du carré pour des valeurs entre 0 et 1

Valeur	Valeur au carré	2 × Valeur	2 × Valeur au carré
0.05	0.0025	0.1	<u>0.005</u>
0.1	<u>0.01</u>	0.2	0.02

Le tableau ci-dessus montre que la somme de deux valeurs basses identiques (0.05) mise au carrée est plus petite que le carré d'une valeur deux fois plus élevée (0.01), tandis que leur simple somme est équivalente. Le but étant de favoriser de manière non linéaire les petites valeurs. L'optimisation de ce paramètre reste à faire.

Il faut noter aussi que la performance d'un ensemble d'ARN à entraîner un algorithme d'apprentissage machine qui a pour tâche de prédire la réactivité chimique des nt. ne dépend pas seulement de ce paramètre.

$$Discriminant = \frac{\left(\sum_{total} \left(1 - abs \left(\frac{o_{Hi}(s-s) - o_{Low}(s-s)}{o_{Hi}(s-s) + o_{Low}(s-s)} \right) \right)^2 \right)}{\quad} \quad \text{Équation 1}$$

Les prochaines pages sont consacrées à des figures représentant la courbe obtenue lorsqu'on classe les S-S par leur ratio de l'équation 1 (Valeur obtenue avant de prendre la valeur absolue). Le discriminant est placé dans l'encadré en haut à gauche. Une autre caractéristique qui indique qu'un ensemble de S-S performera bien, est son nombre de S-S, plus il est grand, meilleur seront les prédictions. Ce nombre est visible à la fin de l'axe horizontal.

Dans les deux figures de droite, trois champs font partie de la légende :

- **Collection** : pour l'ensemble de données défini par les filtres (voir le tableau ci-dessous)
- **Soft** : pour le logiciel de repliement (RNAsubopt ou MCFlashfold)
- **Discriminant** : pour la valeur de l'équation 1 (le plus bas est le mieux)

« Capacité à distinguer » des sous-structures de RNAsubopt

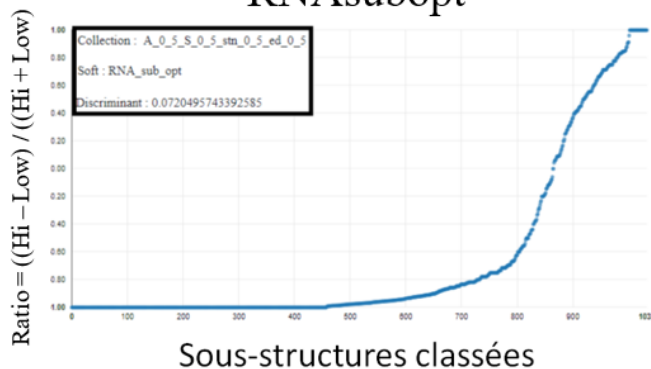


Figure 19. Courbe du pouvoir discriminant des S-S du logiciel de repliement des ARN en SS, RNAsubopt.

« Capacité à distinguer » des sous-structures de MCFlashfold

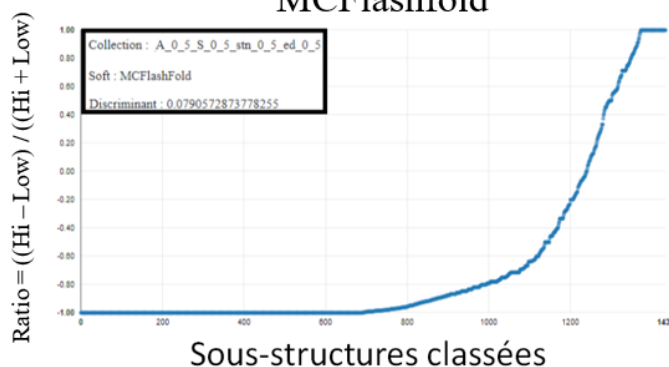


Figure 20. Courbe du pouvoir discriminant des S-S du logiciel de repliement des ARN en SS, MCFlashfold.

On voit dans les deux figures ci-dessus que les courbes se ressemblent beaucoup. Chaque point est une S-S ayant au minimum cinq occurrences dans la base de données (toutes catégories confondues, « Hi », « Bg » et « Low »).

Dans les petits ensembles de données, la valeur du discriminant est plus basse pour le logiciel de RNAsubopt, ce qui indique une meilleure discrimination, mais le nombre

ed

Diversité
d'ensemble
maximum

stn

« Signal to noise »
minimum

S

Score moyen
de réactivité des
nt. maximum

A

Proportion
maximale
d'adénine dans
l'ARN

0_5 = 0,5

de S-S est plus grand chez MCFlashfold, ce qui indique qu'il peut prédire un ensemble plus diversifié d'ARN. L'espace des S-S est aussi moins grand pour MCFlashfold.

Pour des grands ensembles de données, RNAsubopt a de meilleures valeurs dans les deux catégories (voir tableau V pour RNAsubopt et tableau VI pour MCFlashfold).

Choix de l'ensemble d'entraînement

À l'aide des tableaux de la page suivante, les paramètres ont été choisis. Les informations suivantes sont disponibles pour chaque combinaison de paramètres pour les deux logiciels:

1. Le nombre d'ARN de l'ensemble de données (RNA).
2. Le nombre de S-S ayant plus de cinq occurrences (Occ.).
3. Le discriminant (la valeur de l'équation 1, une valeur de 0 est idéale)
4. La valeur de la diversité d'ensemble (DE)

« Capacité à distinguer » des sous-structures de RNAsubopt pour l'ensemble de données qui minimise l'équation 1

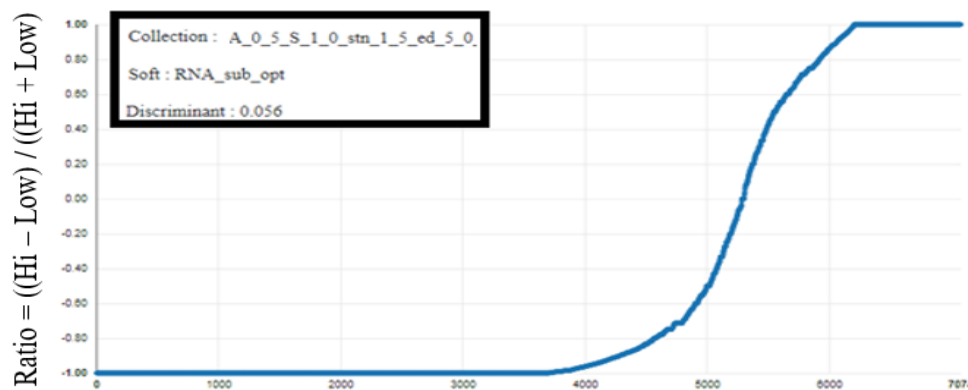


Figure 21. **Courbe du pouvoir discriminant de l'ensemble de données ayant la plus petite valeur pour l'équation 1, le discriminant.** On recherche une pente franche comme celle-ci, moins le nombre de S-S est grand au milieu de l'axe vertical, mieux la classification se fera.

Dans les tableaux de la page suivante, on remarque que le discriminant ne diminue pas avec une augmentation de la DE. Ceci est dû au fait que seule la MFE a été considérée lors de l'obtention des S-S. Ainsi, la réactivité des nt. des ARN ayant une DE élevée est plus difficile à expliquer par les S-S que celle des nt. des ARN ayant un DE basse.

Tableau V. Tableau comparatif des paramètres de l'ensemble d'entraînement du logiciel RNAsubopt

RNAsubopt		Seuil sur le « signal to noise »							
		6				1.5			
		RNA	Occ.	Disc.	DE	RNA	Occ.	Disc.	DE
Seuil sur le score	1	2 715	1 886	0.06	5	18 212	7 074	0.05	5
		4 902	2 915	0.08	20	29 641	11 772	0.07	20
		5 142	3 091	0.09	50	31 713	12 799	0.08	50
	1.5	2 793	1 910	0.06	5	18 492	7 140	0.05	5
		5 037	1 961	0.08	20	30 262	11 972	0.07	20
		5 283	3 139	0.08	50	32 415	13 045	0.08	50
	5	2 793	1 910	0.06	5	18 496	7 140	0.06	5
		5 037	2 961	0.08	20	30 283	11 972	0.07	20
		5 284	3 139	0.09	50	32 440	13 060	0.08	50

Tableau VI. Tableau comparatif des paramètres de l'ensemble d'entraînement du logiciel MCFlashfold.

MCFlashfold		Seuil sur le « signal to noise »							
		6				1.5			
		RNA	Occ.	Disc.	DE	RNA	Occ.	Disc.	DE
Seuil sur le score	1	2 715	2 364	0.12	5	18 212	7 461	0.12	5
		4 902	3 612	0.14	20	29 641	10 751	0.13	20
		5 142	3 779	0.15	50	31 713	11 329	0.14	50
	1.5	2 793	2 415	0.13	5	18 492	7 536	0.12	5
		5 037	3 676	0.14	20	30 262	10 876	0.14	20
		5 283	3 854	0.15	50	32 415	11 471	0.14	50
	5	2 793	2 415	0.13	5	18 496	7 536	0.12	5
		5 037	3 676	0.14	20	30 283	10 876	0.13	20
		5 284	3 854	0.15	50	32 440	11 474	0.14	50

On note aussi que le logiciel RNAsubopt a une valeur de discrimination (Disc.) constamment inférieure au logiciel MCFlashfold. Ceci peut être dû au fait que les SS ont été optimisées avec un logiciel de la même suite que RNAsubopt avant d'être sondées. Dans les tableaux de la page précédente, « Occ. » signifie occurrence des S-S ayant été aperçu plus de cinq fois.

3.3 Prédictions de la réactivité des nucléotides

Lors de la phase de prédiction, chaque nucléotide reçoit une valeur entre -1 et 1. Le signe qualifie la prédiction de vraie ou fautive et un poids entre 0 et 1 module la prédiction. Plus, une S-S se comporte de manière constante quant à la réactivité de ses nt., plus le score du poids se rapproche de 1. Contrairement, à un score fixe, ce poids permet de diminuer l'impact d'une prédiction pour laquelle les données ne sont pas en accord. De plus, il permet d'augmenter la pénalité d'une prédiction se trompant grandement. RNASS_v2 produit deux informations distinctes par nt. qui peuvent être combinés pour donner une valeur qui varie entre -1 et 1.

Lorsqu'une S-S a un nombre significatif de nt. appartenant à la classe de réactivité basse ou haute, la prédiction est considérée comme fiable et l'algorithme peut se prononcer. La somme des valeurs de prédictions modulée par le poids des nt. de chaque conformation donne une idée du niveau de confiance que nous avons pour cette conformation.

Les S-S des nt. qui n'ont pas assez de données et les nt. qui n'ont pas de valeur du tout n'ont pas de prédiction. L'étape de la prédiction est l'étape limitante dans l'algorithme. Ceci est dû au grand nombre de requêtes fait à la base de données. Ce nombre est égal au nombre de SS multiplié par le nombre de nt. de la séquence pour chaque logiciel de prédiction par ARN. Le calcul de seulement 10 SS de 1000 ARN ayant 100 nt. pour lesquels on a une valeur de réactivité demande 1 million de requêtes à la base de données, d'où l'importance d'optimiser les performances de la base de données en compilant les S-S par valeur de réactivité et en les indexant sur les S-S et les logiciels.

Conclusion du chapitre 1

RNASS est un ensemble de fonctions intégrant trois logiciels de prédiction des SS. Il crée des fichiers standardisés ayant une structure arborescente. Ces fichiers sont dans le format JSON. Vu la structure de ces fichiers, il est facile d'ajouter des informations sans compromettre leur compatibilité avec les logiciels qui les utilisent. Toutes les données qu'ils contiennent pourraient être calculées à la volée, mais leur « précalculé » augmente la fluidité des outils d'analyses qui leur sont dédiés. Les limites du « précalculé » sont surtout l'espace disponible et la connaissance des besoins futurs. Cela étant dit, l'espace mémoire est de moins en moins coûteux, cela rend le « précalculé » de plus en plus avantageux.

Le prochain chapitre présente RDV un logiciel qui permet de visualiser les données de RNASS_v2.

Chapitre 2 : RDV : Visualisation de la SS des ARN

RDV est l'abréviation de RNA Dynamic Viewer. Ce logiciel permet de visualiser les données provenant de RNASS_v2. Il a été écrit en *Javascript / Type script* et utilise *MongoDB, Express, Angular* et *Node.js*. Cette suite de logiciels est nommée communément le « *MEAN stack* » une suite qui permet de créer des sites internet modernes. RDV superpose les données de réactivités chimiques sur les SS d'un ARN. Tout comme le logiciel de visualisation de SS FORNA, RDV utilise une librairie codée en *Javascript* nommé *D3.js* pour représenter la SS d'un ARN. Cette librairie implémente un champ de force dont l'objectif est de dessiner des réseaux formés de nœuds et de liens.

4.1 Visualisation du graphe des transitions

Parmi les fonctions innovantes de RDV se trouve la possibilité de voir un ensemble de SS les unes après les autres en continuité. Il est ainsi plus facile de voir la différence entre deux SS.

La transition fluide entre les SS permet aussi de mieux comprendre la dynamique des ARN. Grâce à ce logiciel, il est possible d'avoir une vue d'ensemble sur les ARN sondés. L'avantage de la flexibilité des composants est la possibilité de placer l'ARN de façon à faire ressortir une information importante.

Dans la vue principale, les réseaux de transition des premières SS de chaque logiciel sont dessinés à la droite de la SS actuelle. Chaque cercle représente une SS et leur position sur l'axe vertical est proportionnelle à leur énergie, plus l'énergie des nt. est basse plus le cercle sera bas. Un lien est dessiné entre deux SS partageant 90 % de leurs paires de bases. Un « click » sur un SS, la sélectionne.

4.2 Visualisation de la cohérence

Une autre fonction innovante de RDV est la possibilité de valider une SS prédite en se basant sur la réactivité observée de ses S-S.

La couleur des SS dans le graphe des transitions est reliée à la vraisemblance de la SS, plus elle est pâle meilleur elle est.

Le contour des nt. est relié à la prédiction de RNASS_v2. Un contour rouge signifie une réactivité moyenne basse et un contour bleu signifie une réactivité moyenne élevée. L'épaisseur du contour est reliée à la confiance de la prédiction, ce qui fait ressortir les mauvaises prédictions ayant une confiance élevée. La valeur exacte de la prédiction globale est disponible en plaçant la souris au-dessus de la conformation ou du nucléotide d'intérêt.

4.3 Visualisation de la SS

La SS sélectionnée est dessinée du côté gauche de la visualisation. Dans cette visualisation, chaque cercle correspond à un nucléotide, trois modes de couleur sont disponibles,

- le mode : couleur selon la sorte du nucléotide
- le mode : couleur selon la réactivité du nucléotide.
- le mode : couleur selon l'erreur reliée à chaque nucléotide.

4.4 Obtenir des détails et rechercher des ARN semblables

La nécessité d'avoir un outil de visualisation pour ce type de prédiction vient de la présence de nt. se comportant de façon inattendue. Ces nt. sont extrêmement intéressants lorsque vient le temps de comprendre les interactions tertiaires des nt. au

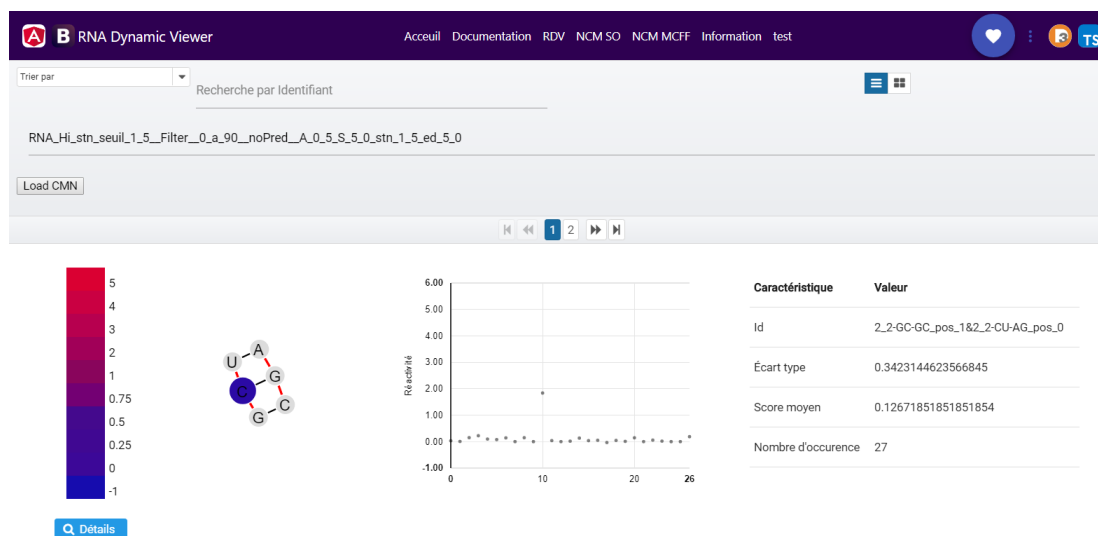


Figure 22. **Vue détaillée sur les S-S dans RDV.** Une échelle de couleur permet d'avoir une idée rapide de la réactivité moyenne des nt. de la S-S dessinée à sa droite. Les points du graphique au centre représentent chacune des occurrences de la S-S. L'échelle à gauche est graduée de -1 à 6. Une valeur élevée au-dessus de 1 représente un nt. réactif tandis qu'une valeur près de 0 indique une absence de réactivité. L'identifiant, le score moyen, l'écart-type et le nombre d'occurrences de la S-S sont complètement à droite de la figure. Lorsqu'on clique sur le bouton « Détails », une chaîne de caractère avec toutes les valeurs de réactivité s'affiche. En passant la souris sur chaque point, on obtient l'expérience d'où provient le nucléotide et son numéro d'ARN. Il est possible de faire une recherche par identifiant et de charger différents ensembles des données.

sein d'un ARN, mais sont difficilement explicables par les S-S. Heureusement, la majorité des interactions sont locales.

Dans la vue globale, voir figure 23, la couleur des nœuds du réseau des conformations est reliée à cette somme. Le gradient de couleur passe du vert pâle au vert foncé pour le logiciel MCFlashfold et du rouge pâle au rouge foncé pour le logiciel RNAsubopt. Dans les deux cas, une couleur pâle signifie une moins bonne prédiction qu'une couleur foncée.

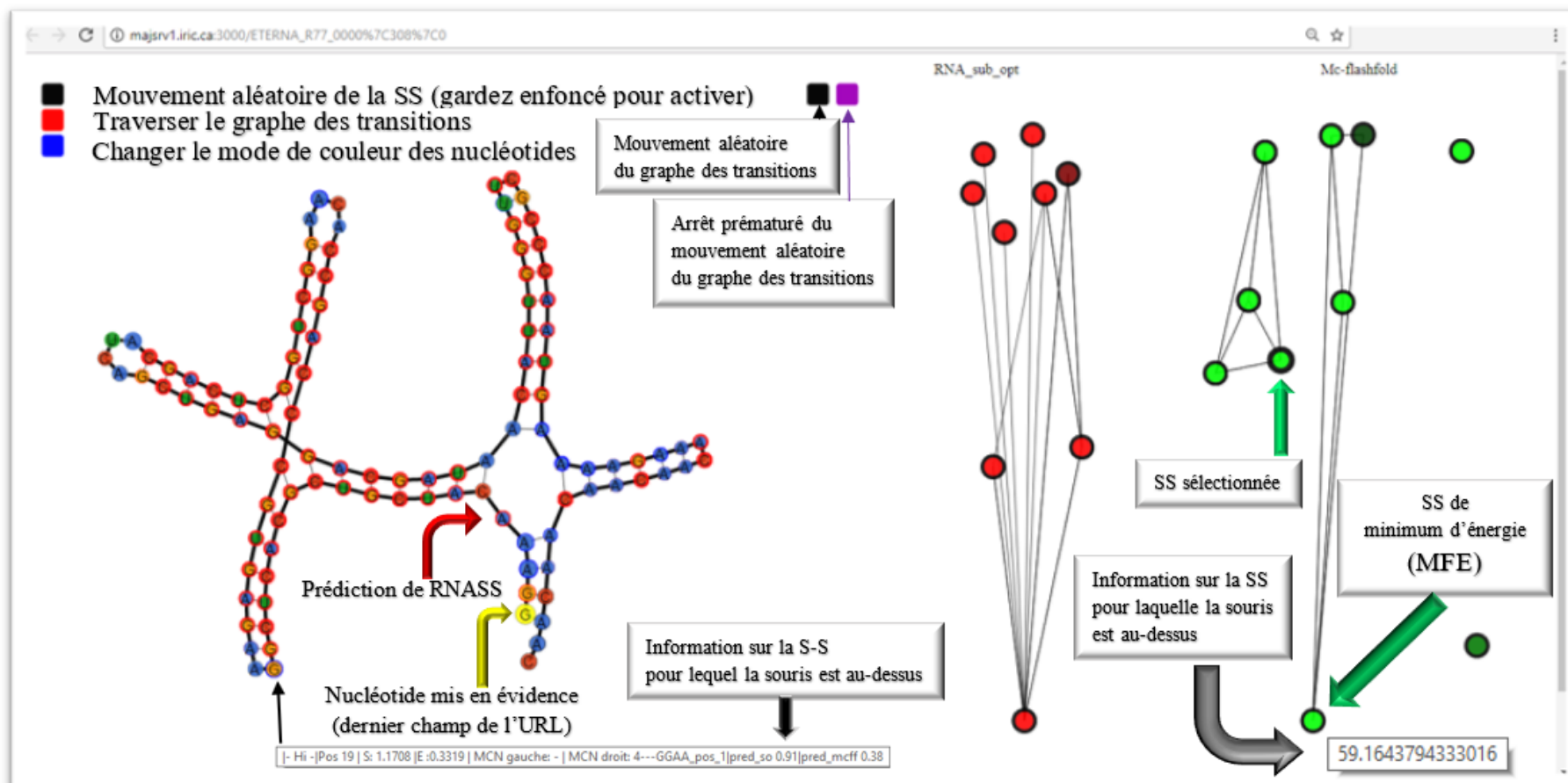


Figure 24. **Vue principale de RNA Dynamic Viewer (RDV).** La « simulation » commence en spirale et se déploie automatiquement. Pour agiter les nt., on clique sur le rectangle noir en haut à gauche. Celui qui est **rouge** lance et arrête la traversée des graphes de transitions. On change le mode de couleur en appuyant sur le rectangle **bleu**. En cliquant sur le cercle lui correspondant dans le graphe des transitions, on change la SS. Les performances de la simulation dépendent des performances de votre ordinateur, il a été testé dans le navigateur « *chrome* ». L'objet « *jsonData* » nous renseigne sur l'ARN dans la console JavaScript (f12).

Conclusion du chapitre 2

RDV a deux vues principales complémentaires. La première est la vue locale des S-S. Cette vue permet d'avoir une idée de la réactivité des nucléotides dans le contexte de leur S-S. Par cette vue, nous sommes en mesure d'identifier les ARN dans lesquels la S-S est présente. La deuxième vue, est la vue globale dans laquelle on voit tout l'ARN replié en SS avec en superposition la réactivité des nt.. Les 10 premières SS sont représentées par des cercles dont la hauteur est proportionnelle à leur énergie. La MFE étant placée dans le bas de l'image.

L'outil de visualisation RDV se démarque des autres logiciels de visualisation de SS par la possibilité de voir le graphe des transitions des ARN. Une des caractéristiques innovantes, non exploitées dans ce projet de recherche, est qu'il permet la comparaison des SS basée sur la somme des valeurs de cohérence de chaque nt.. La valeur de cohérence est calculée en prenant en compte la prédiction faite par RNASS_v2 et la valeur réelle de réactivité.

J'ai placé à la toute fin de l'annexe plusieurs figures qui montrent le potentiel de ce logiciel. Par exemple, il est possible de faire le portrait de la réactivité d'une boucle, nt. par nt.. De plus, grâce au graphique interactif montrant la réactivité de chacun des nt. (de la vue locale des S-S) nous pouvons approfondir nos recherches sur les nt. qui ne réagissent pas comme les autres. En outre, ce logiciel nous permet de voir qu'il serait intéressant d'explorer le lien entre les SS sous-optimales et les nt. difficiles à prédire. Finalement, je crois que cet outil a un avenir intéressant devant lui puisqu'il est écrit en *JavaScript* avec la librairie d3.js, une librairie libre.

Le code de RDV est libre lui aussi. Voir : <https://github.com/PhilippeMalric/rdv.git>

Chapitre 3 : Évaluation du modèle des cycles simples

Dans le but de mesurer les performances des prédictions effectuées par RNASS, je les ai évalués dans ce chapitre. J'ai utilisé une machine à vecteur de support pour entraîner un modèle avec deux ensembles de données différents.

Le premier ensemble contient les valeurs de pairage des nt. et le deuxième contient les valeurs de prédiction de RNASS.

Le graphique de la figure 24 montre les étapes de l'entraînement. La plateforme Azure de Microsoft ® permet de créer un protocole graphique.

En partant du haut de la figure 24 vers le bas, voici une description des étapes (les icônes réfèrent aux icônes de la figure 24 (voir page suivante) :



1. Les données sont dans cet élément. 2 116 180 nt. sont considérés. Ils ont une valeur de réactivité haute (>1) ou basse ($<0,5$).



2. On configure les métadonnées pour que l'algorithme d'apprentissage sache sur quelles caractéristiques apprendre et quelle caractéristique prédire.



3. On choisit le modèle d'apprentissage et on lui donne les paramètres voulus. Dans le présent cas, les paramètres sont ceux d'origines.



4. On sépare les données pour avoir un ensemble d'entraînement et un ensemble « test ». Ici, on a séparé en deux l'ensemble de départ.



5. On entraîne le modèle.



6. On attribue un score avec l'ensemble « test ».



7. On l'évalue. Dans cette étape on calcul les performances avec les métriques expliquées dans l'introduction.

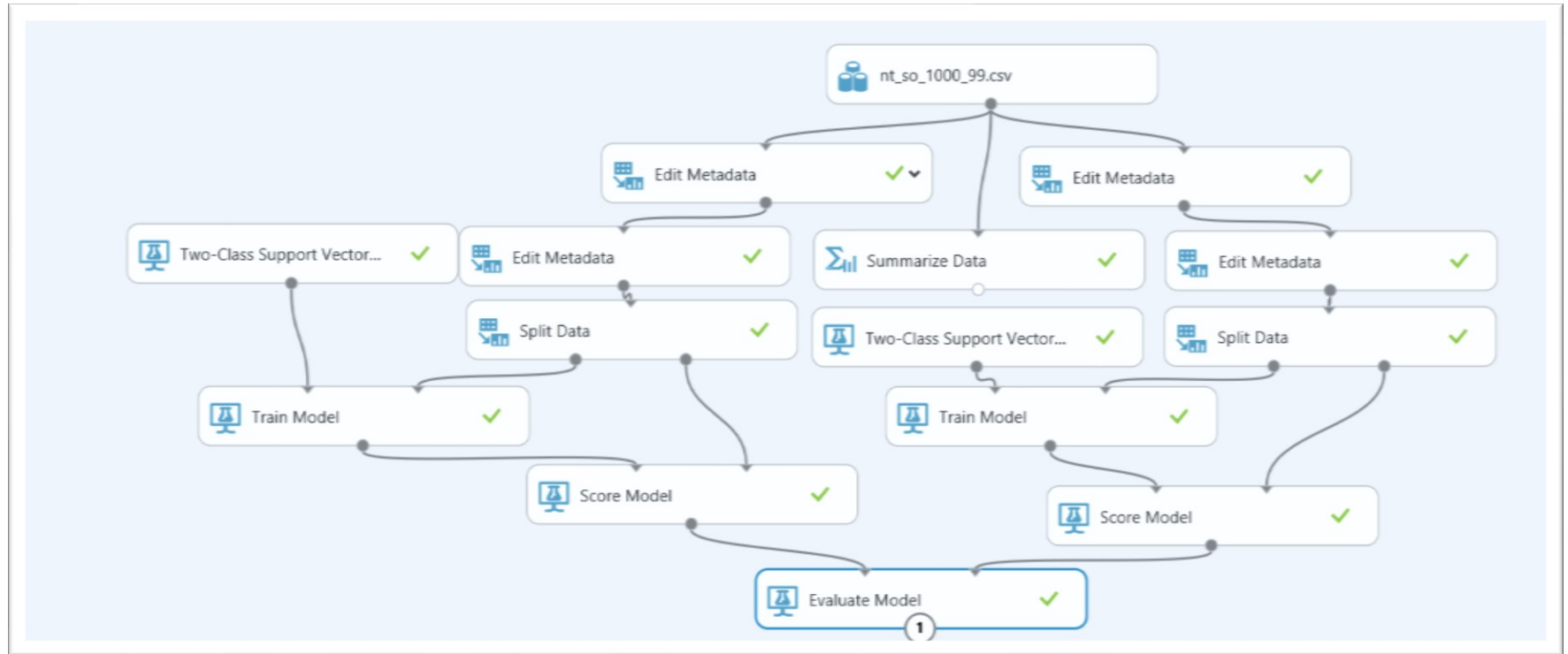


Figure 25. **Protocole d'apprentissage graphique, partant des données brutes jusqu'à leur analyse.** La plateforme Azure permet d'entraîner un modèle en effectuant de la programmation graphique. Si on fait abstraction du bloc « *summarise data* », deux chemins partent des données brutes. À gauche, le modèle est entraîné sur les prédictions de RNASS et à droite sur l'état pairé ou non des nt.. Dans les deux cas, une machine à vecteur de support a été utilisée. Le dernier bloc rassemble les données pour créer des graphiques, tels que la courbe ROC et la précision en fonction du rappel.

5.1 Comparaison des prédictions de RNASS avec celles faites à l'aide de l'état pairé ou non d'un nucléotide.

Les modèles ont été entraînés et évalués sur deux ensembles distincts de plus d'un million de nt. (1 058 090) provenant des prédictions de la SS de RNAsubopt. Pour les deux modèles, deux ensembles ont été créés en séparant de façon aléatoire un grand ensemble de 2 116 180 nt., l'ensemble d'entraînement et l'ensemble « test ».

Tableau VII. Tableau de contingence de l'état pairé ou non des nt. en fonction de leur niveau de réactivité.

Total : 1 058 090 nucléotides		État des nucléotides	
		Non pairés	Pairés
Réactivité	Haute	155 247	22 179
	Basse	210 901	669 763

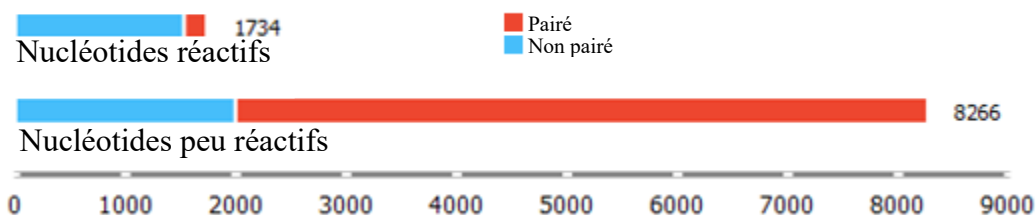


Figure 26. **Table de contingence de l'état (pairé ou non) des nt. avec leur réactivité chimique.** Le nombre de nt. réactifs non pairés est semblable au nombre de nt. peu réactifs non pairés. Cependant, il y a beaucoup moins de nt. réactifs pairés que de nt. peu réactifs pairés. Cette figure a été faite en prenant en compte 10 000 nt..

Le pourcentage de nt. pairés dans ces ensembles est d'environ 65 % et le pourcentage de nt. peu réactifs est de 83 %. Ces deux pourcentages supposent déjà que certains nt. peu réactifs seront non pairés. En réalité, c'est 20% des nt. totaux qui sont peu réactifs et non pairés. Parmi les nt. peu réactifs, seulement 24% ne sont pas pairés. En revanche, 12.5 % des nt. avec une réactivité élevée sont pairés. Ces pourcentages relativement élevés sont le signe que l'état pairé ou non n'explique pas bien la réactivité des nt..

La figure 26 est une courbe ROC des deux modèles superposés. De gauche à droite, le taux de faux « Hi » augmente linéairement. Plus le taux de faux « Hi » est bas pendant que le taux de vrai « Hi » est élevé, meilleur est l'algorithme.

RNASS donne une probabilité avec ses prédictions. Dans le cas du modèle basé sur l'état pairé ou non des nt., le nombre de prédictions « Hi » est fixe, c'est la raison de la forme « carrée » de la « courbe » rouge du graphique de la figure 24. En fait, un seul point est calculé. On constate que le modèle basé sur l'état des nt. à un taux de faux « Hi » d'environ 24 %, ce qui correspond à un taux de vrai « Hi » de 87.4 %. Pour un même taux de faux positifs, RNASS à un taux de vrai « Hi » au-dessus de 90 %, un gain de plus de 2.4 %

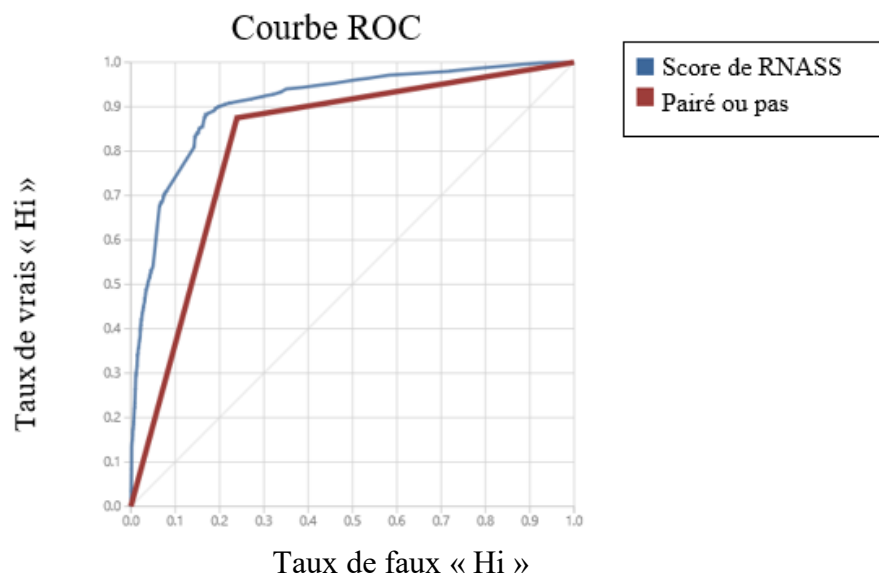


Figure 27. **Courbe de ROC du modèle de RNASS et du modèle pairé / non pairé.** Le modèle de RNASS en plus d'être paramétrable au niveau du risque de ses prédictions performe mieux que le modèle basé sur l'état pairé ou non des nt..

Lorsqu'on considère les nt. réactifs comme des positifs, la précision du modèle pairé / non pairé est de 42.4 % (la précision lorsqu'on considère les nt. peu réactifs comme positifs est de près de 97 %).

Cela veut dire que 42.4 % des nt. non pairés sont réactifs (calcul effectué avec les valeurs de la figure 42). Pour le même nombre de vrai « Hi », RNASS obtient une précision de 51 %. Cependant, le modèle de RNASS peut être ajusté. Par exemple, si l'on désire une précision de 85 % sur nos prédictions « Hi », on le peut, mais le rappel est de seulement 2 % (voir le graphique précision en fonction du rappel de la figure ci-dessous).

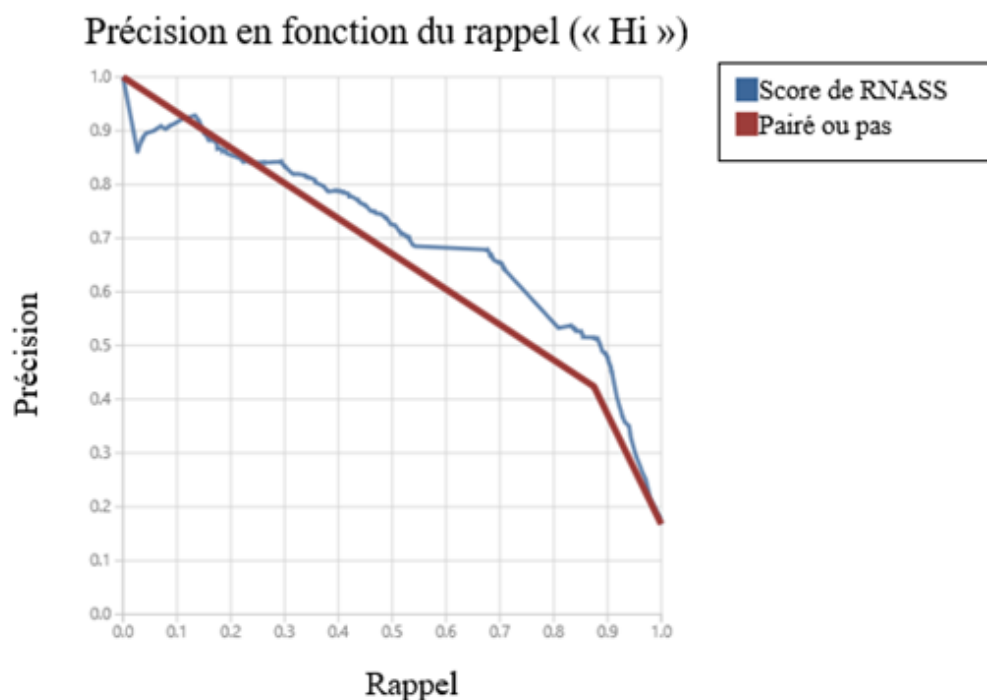


Figure 28. **Précision en fonction du rappel du modèle de RNASS (en bleu) et de celui basé sur l'état pairé non pairé des nt. (en rouge).** Plus de 1 million de nt. ont servis à faire ce test. Le modèle de RNASS performe mieux que celui basé sur l'état pairé ou non des nt..

Chose surprenante, du point de vue du nombre des bonnes prédictions total divisé par le nombre de prédictions total, pour le modèle prenant seulement en compte l'état pairé ou non d'un nt ; il est plus avantageux de prédire tous les nt. comme étant de basse réactivité plutôt que de se fier sur le pairage de ceux-ci. Autrement dit, le nombre de nt. non pairés et peu réactifs est tellement

grand par rapport au nombre de nts non pairés et réactifs que l'état non pairé aide peu à trouver les nt. réactifs (moins de un nt. non pairé sur deux est réactif).

Tableau VIII. Performance des algorithmes d'apprentissage machines à prédire la réactivité chimique des nt..

Méthodes	Métriques				
	AUC	Accuracy	F1 Score	Précision	Rappel
AdaBoost	0.924	0.886	0.703	0.702	0.704
C. bayésien naïf	0.899	0.88	0.703	0.673	0.736
Réseaux de neurones	0.916	0.88	0.697	0.677	0.718
Regression Logistique	0.909	0.88	0.691	0.683	0.7
Random Forest	0.924	0.886	0.702	0.703	0.702

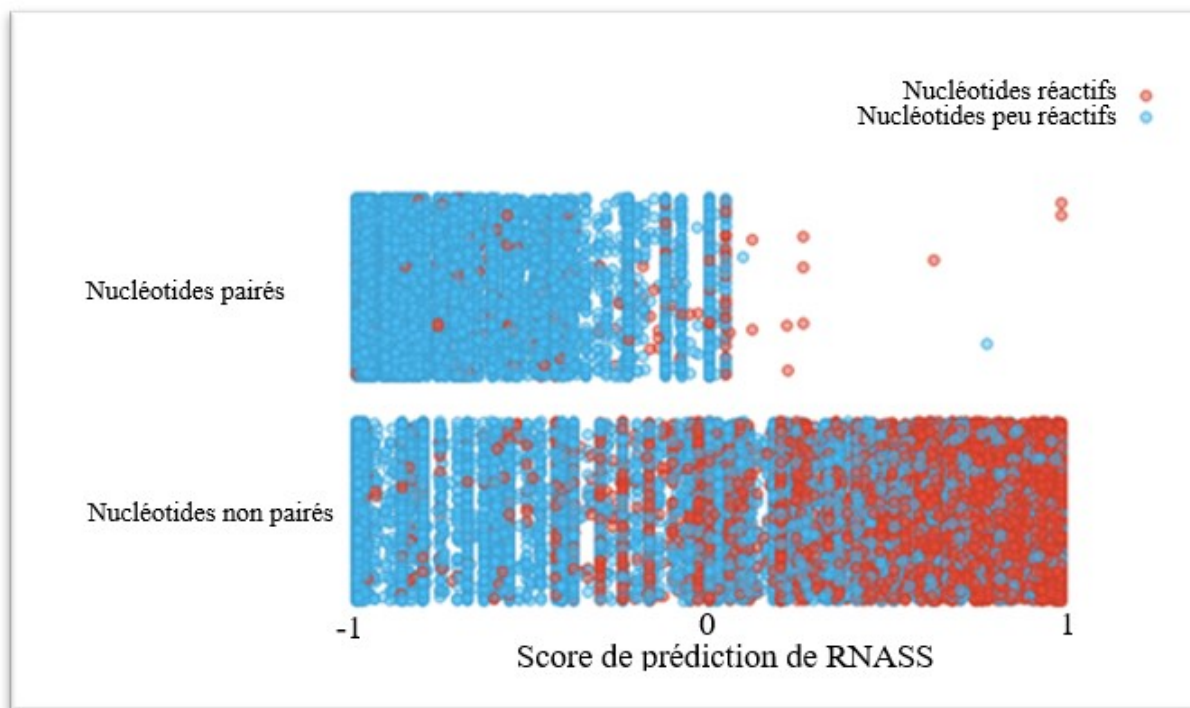


Figure 29. Distribution des nt. en fonction de leur état (axe vertical), de leur score de prédiction de RNASS (axe horizontal) et de leur réactivité chimique (couleur). Les points rouges représentent les nt. réactifs. Ils sont beaucoup plus présents lorsque le nucléotide est non pairé et lorsque son score de prédiction est élevé. Cette visualisation a été faite avec Orange, un logiciel d'apprentissage machine.

Le tableau IV montre quelques valeurs des métriques de performance du modèle basé sur le score de RNASS. Pour obtenir un taux de bonnes prédictions de plus de 90% sur les nt. de haute réactivité, il faut diminuer le rappel en dessous du 1 %. Les nt. de la figure 28 ont été divisés en deux groupes, les nt. pairés en haut et ceux non pairés en bas. Les nt. sont ordonnés selon leur score de prédiction, -1 à gauche jusqu'à 1 à droite. La hauteur des nt. n'a pas d'importance. Les nt. rouges sont ceux qui ont un score de réactivité au-dessus de 1 et ceux en bleu ont un score de réactivité en dessous de 0.5.

Conclusion du chapitre 3

Le score de prédiction de RNASS performe mieux que les prédictions faites avec l'état pairé ou non des nt.. Il se démarque par le fait qu'il peut prendre plus ou moins de risque dépendamment des besoins. Par exemple, lorsqu'on cherche un nt. réactif dans un ARN.

Au fur et à mesure que d'autres données sur les ARN seront recueillies, ce score pourra être amélioré. En effet, pour certaines S-S le nombre de nt. est trop bas pour émettre des bons jugements. Puisque la qualité des prédictions des SS utilisés lors de l'entraînement de ce score est cruciale à son fonctionnement des améliorations de ce côté ne peuvent qu'aider le modèle.

Dans ce chapitre, les S-S utilisées étaient des cycles simples. De nouvelles données montrent que les S-S composées de plusieurs cycles performant encore mieux. Aussi, il serait possible d'inclure l'information donnée par les SS sous-optimales dans un prochain modèle. La difficulté posée par l'ajout de ces informations additionnelles est qu'elles augmentent la grandeur de l'espace de recherche, ce qui diminue le nombre d'occurrence d'un cas en particulier. Heureusement, le nombre de séquences d'ARN sondées augmente lui aussi d'année en année.

Conclusion globale

Les règles permettant de prédire la réactivité d'un nucléotide sont beaucoup plus complexes que le simple fait qu'un nucléotide soit pairé ou non. Lorsqu'on considère seulement l'état pairé ou non des nts, les nts ayant une basse réactivité sont plus faciles à prédire que ceux ayant une réactivité élevée parce qu'il y en a beaucoup plus et que certaines S-S ont une réactivité très stable (les paires de bases Watson-Crick (C-G/ A-U)) dans une hélice par exemple).

On doit trouver des règles plus performantes que le non-appariement des nt. pour expliquer leur réactivité. Les règles proposées dans ce mémoire sont basées sur la structure locale du nucléotide, les cycles et les S-S. Elles pourraient être raffinées et simplifiées pour les généraliser.

Le taux élevé de faux « Hi » nous amènent à faire de fausses prédictions lorsqu'on veut prédire la ou les SS d'un ARN en particulier. En prenant en compte les S-S cette erreur diminue. Il faut noter que l'information du pairage d'un nucléotide est inscrite dans la S-S ce qui lui assure de mieux performer sauf dans les cas où les données sur les S-S sont rares.

Statistiquement, plus une S-S est petite, plus il y aura d'occurrences de celle-ci dans la base de données et moins le nombre total de S-S sera grand. Ceci favorise les logiciels prédisant plusieurs petits cycles au lieu de quelques grands. MCFlashfold répond à ces exigences.

Avec mon approche, les ARN ont un score de cohérence basé sur la correspondance entre le niveau de réactivité de leur S-S et le niveau de réactivité observé.

Mon algorithme s'appuie sur le fait que les logiciels de prédiction de SS font de bonnes prédictions en moyenne (lorsqu'on sélectionne bien les ARN). En réalité, il ne fait que vérifier la cohérence des logiciels avec eux même et lève des drapeaux lorsque ce n'est pas le cas.

Champ d'études à venir et algorithmes à considérer

Les liens entre les nt. distants sont difficiles à prédire, mais nous savons qu'ils existent, il est donc essentiel de laisser une marge de manœuvre aux modèles. Cette liberté laisse la place à des hypothèses qui pourront être réfutées ou confirmées. Dans un article d'Alain Denise[39], les liens tertiaires sont calculés en minimisant une fonction de coût qui oblige la majorité des boucles à faire une interaction avec un autre élément de l'ARN. Ce groupe a obtenu de bons résultats et leur méthode pourrait être appliquée à ce modèle ou vice et versa. Précisément, la non-réactivité anormale d'une boucle pourrait se voir attribuer une importance plus grande dans la fonction de coût, ce qui l'obligerait à être pairé.

Pour le moment aucun logiciel offert librement ne montre la dynamique d'un ARN à partir d'une SS. RDV permet d'incorporer cette fonction. La modification du champ de force en y ajoutant des contraintes physiques par d'autres moyens tels que la RMN combinée à une visualisation de la SS en 3D, nous donnera une meilleure idée des mouvements des molécules. Très probablement, des hypothèses naîtront de cette visualisation. Le mécanisme des ARN catalysant des réactions chimiques serait un bon exemple à étudier.

Mon analyse s'inspire du processus scientifique. Les ARN et les SS prédites entrant dans la construction de la base de données des cycles sont fiables en raison des contrôles effectués et de leur nature même. Dans un prochain projet, il sera pertinent d'élargir l'espace de recherche, dans notre cas, apprendre des séquences plus difficiles à prédire. On pourrait par exemple :

- abaisser le rapport signal sur bruit
- augmenter la diversité d'ensemble
- augmenter le score moyen de réactivité des ARN.

Pour mon algorithme, ce qui détermine si une séquence est difficile à prédire ou cohérente, c'est la combinaison de sa composition en S-S et les valeurs de réactivité associées. Plus le nombre de S-S ayant une réactivité variable (autant de « Hi » que de « Low ») est grand, plus l'ARN est considéré comme difficile et obtient un score global bas. Ces ARN ont des mystères à élucider

et beaucoup de questions restent sans réponse encore à ce jour. Par exemple, un ARN peut-il alterner entre deux SS rapidement comme on le voit dans la simulation de RDV?

Dans le futur, les expériences de réactivité chimique seront surement guidées par des algorithmes d'apprentissage machine. Un bon algorithme d'apprentissage par rétroaction devrait garder le nombre de S-S variables constamment petit, mais non nul. Ainsi, l'espace total des incertitudes se rétrécirait.

Ce processus itératif tirerait avantage des « mauvaises prédictions », il les transformerait en occasions d'apprentissage. Ultimement, cet algorithme convergera vers un ensemble de règles qui expliquerait tous les phénomènes de réactivités chimiques des ARN, mais pour atteindre ce but beaucoup de sondages chimiques des ARN restent à faire. Orienter ces expériences a été l'un des objectifs de mon approche. Les S-S ayant une faible occurrence et ceux variables doivent être plus étudiés.

Finalement, les interactions tertiaires devraient être prises en compte, un algorithme qui listerait les S-S distantes entraînant une réactivité anormale pour chaque S-S variable promet de conduire à de nouvelles découvertes.

Bibliographie

1. Vangaveti, S., S.V. Ranganathan, and A.A. Chen, *Advances in RNA molecular dynamics: a simulator's guide to RNA force fields*. Wiley Interdisciplinary Reviews: RNA, 2017. **8**.
2. Deigan, K.E., et al., *Accurate SHAPE-directed RNA structure determination*. Proceedings of the National Academy of Sciences, 2009. **106**: p. 97-102.
3. Hajdin, C.E., et al., *Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots*. Proceedings of the National Academy of Sciences, 2013. **110**: p. 5498-5503.
4. Lucks, J.B., et al., *Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)*. Proceedings of the National Academy of Sciences, 2011. **108**: p. 11063-11068.
5. Yesselman, J.D., et al., *Updates to the RNA mapping database (RMDb), version 2*. Nucleic Acids Research, 2017: p. 1-5.
6. Mandal, M. and R.R. Breaker, *Gene regulation by riboswitches*. Nature reviews. Molecular cell biology, 2004. **5**: p. 451-463.
7. Almeida, M.I., et al., *MicroRNAs: genomics, biogenesis, mechanism, and function*. Cell, 2015. **33**: p. 4663-4670.
8. Weill, N., et al., *MiRBooking simulates the stoichiometric mode of action of microRNAs*. Nucleic Acids Res, 2015. **43**(14): p. 6730-8.
9. Dahm, R., *Friedrich Miescher and the discovery of DNA*. Developmental Biology, 2005. **278**: p. 274-288.
10. Caspersson, T. and J. Schultz, *Pentose Nucleotides in the Cytoplasm of Growing Tissues*. Nature, 1939. **143**: p. 602.
11. Watson, J.D. and F.H.C. Crick, *Molecular structure of nucleic acids*, in *Nature*. 1953. p. 737-738.
12. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins*. Journal of Molecular Biology, 1961. **3**(3): p. 318-356.
13. Holley Rw Fau - Apgar, J., et al., *STRUCTURE OF A RIBONUCLEIC ACID*. (0036-8075 (Print)).
14. Tinoco I Jr Fau - Uhlenbeck, O.C., M.D. Uhlenbeck Oc Fau - Levine, and M.D. Levine, *Estimation of secondary structure in ribonucleic acids*. (0028-0836 (Print)).
15. Nussinov, R., et al., *Algorithms for Loop Matchings*. SIAM Journal on Applied Mathematics, 1978. **35**(1): p. 68-82.
16. Brassard, G. and P. Bratley, *Fundamentals of Algorithmics*. 1995. p. 524.
17. Zuker, M., *On finding all suboptimal foldings of an RNA molecule*. 1989(0036-8075 (Print)).
18. Zuker, M. and D. Sankoff, *RNA secondary structures and their prediction*. Bulletin of Mathematical Biology, 1984. **46**(4): p. 591-621.
19. McCaskill, J.S., *The equilibrium partition function and base pair binding probabilities for RNA secondary structure*. (0006-3525 (Print)).

20. Major, F., et al., *The combination of symbolic and numerical computation for three-dimensional modeling of RNA*. Science (New York, N.Y.), 1991. **253**: p. 1255-60.
21. Hofacker, I.L., et al., *Fast folding and comparison of RNA secondary structures*. Monatshefte für Chemie Chemical Monthly, 1989. **125**: p. 167-188.
22. Serra, M.J. and D.H. Turner, *Predicting thermodynamic properties of RNA*. (0076-6879 (Print)).
23. Parisien, M. and F. Major, *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data*. Nature, 2008. **452**: p. 51-55.
24. Patel, D.J., *Structural analysis of nucleic acid aptamers*. Current opinion in chemical biology, 1997. **1**: p. 32-46.
25. Bolton, E.E., et al., *PubChem3D: A new resource for scientists*. Journal of Cheminformatics, 2011. **3**: p. 1-15.
26. Staple, D.W. and S.E. Butcher, *Pseudoknots: RNA structures with diverse functions*. PLoS Biology, 2005. **3**: p. 0956-0959.
27. Vendeix, F.A., A.M. Munoz, and P.F. Agris, *Free energy calculation of modified base-pair formation in explicit solvent: A predictive model*. RNA, 2009. **15**(12): p. 2278-87.
28. Zuker, M. and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information*. Nucleic Acids Research, 1981. **9**: p. 133-148.
29. Weeks, K.M. and D.M. Mauger, *Exploring RNA structural codes with SHAPE chemistry*. 2012. **44**: p. 1280-1291.
30. Mathews, D.H., *Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization*. Rna, 2004. **10**: p. 1178-1190.
31. Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure*. Proceedings of the National Academy of Sciences, 2004. **101**: p. 7287-7292.
32. Lorenz, R., I.L. Hofacker, and P.F. Stadler, *RNA folding with hard and soft constraints*. Algorithms for Molecular Biology, 2016. **11**: p. 8.
33. Kladwang, W., et al., *Standardization of RNA chemical mapping experiments*. Biochemistry, 2014. **53**: p. 3063-3065.
34. Seetin, M.G., et al., *Massively Parallel RNA Chemical Mapping with a Reduced Bias MAP-Seq Protocol*. 2014. **1086**: p. 95-117.
35. Lucks, J.B., et al., *Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)*. Proc Natl Acad Sci U S A, 2011. **108**(27): p. 11063-8.
36. Kasner, E., et al., *The Mechanisms of RNA SHAPE Chemistry*. 2015. **70**: p. 646-656.
37. Mlynsky, V. and G. Bussi, *Molecular Simulations Reveal an Interplay Between SHAPE Reagent Binding and RNA Flexibility*. The Journal of Physical Chemistry Letters, 2017: p. acs.jpcl.7b02921.
38. d'Eterna, J.
39. Lamiable, A., et al., *An Algorithmic Game-Theory Approach for Coarse-Grain Prediction of RNA 3D Structure*. 2013. **10**: p. 193-199.
40. Lorenz, R., et al., *ViennaRNA Package 2.0*. Algorithms for Molecular Biology, 2011. **6**: p. 26.

41. Gorodkin, J. and J.M. Walker, *RNA Sequence , Structure , and Function : Computational and Bioinformatic Methods IN Series Editor*.
42. Dallaire, P., *Une signature du polymorphisme structural d ' acides ribonucléiques non-codants permettant de comparer leurs niveaux d ' activités biochimiques par Résumé*. 2014.
43. Flamm, C., et al., *Barrier Trees of Degenerate Landscapes*. Zeitschrift für Physikalische Chemie, 2002. **216**: p. 155.
44. Lee, J., et al., *RNA design rules from a massive open laboratory*. Proceedings of the National Academy of Sciences, 2014. **111**: p. 2122-2127.
45. Das, R., *Eterna Forum*. 2014.
46. Darty, K., A. Denise, and Y. Ponty, *VARNA: Interactive drawing and editing of the RNA secondary structure*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010. **7**: p. 309-322.
47. Kerpedjiev, P., S. Hammer, and I.L. Hofacker, *Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams*. Bioinformatics, 2015. **31**: p. 3377-3379.
48. leinbaum, D.G., M. Klein, and E.R. Pryor, *Logistic regression: a self-learning text*. 2002.
49. Steinwart, I. and A. Christmann, *Support Vector Machines*. 2008: p. 618.
50. Barros, R.C., A.C.P.L.F. de Carvalho, and A.A. Freitas, *Automatic Design of Decision-Tree Induction Algorithms*. 2015.
51. Friedman, J.H., *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 2001. **29**: p. 1189-1232.
52. Herbrich RHERB, R., et al., *Bayes Point Machines*. Journal of Machine Learning Research, 2001. **1**: p. 245-279.
53. POWERS, D.M.W., *Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies, 2011. **2**: p. 37-63.
54. Watkins, A., et al.; Available from: <https://rmdb.stanford.edu/deposit/specs/>.

Annexe

Caractéristiques du serveur utilisé

24 cores, 96GB : Intel(R) Xeon(R) CPU X5650 @ 2.67GHz

API d'Eterna

L'université Stanford rend disponible plusieurs ensembles de données par une interface de programmation (API) distincte de la RMDB, ils ont tous un identifiant sous forme de nombre. Pour obtenir les données brutes en format tsv (tab separated value), on remplace « !id » par l'identifiant dans cette URL :

- <http://www.eternagame.org/tsv/synthesis!id.tsv>

On peut aussi explorer les données avec l'interface d'Eterna en utilisant cette URL:

- <http://www.eternagame.org/game/browse!/id>

Pour retrouver notre ARN d'intérêt, on peut le chercher avec sa séquence.

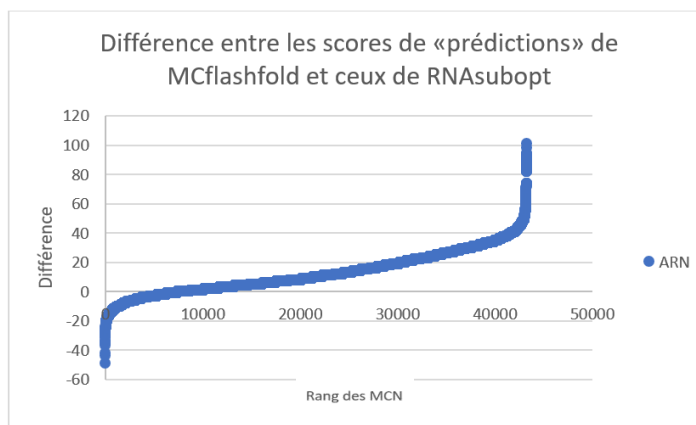
L'ensemble des identifiants des laboratoires est disponible à cette adresse :

<http://www.eternagame.org/get/?type=labs&size=400&skip=0>

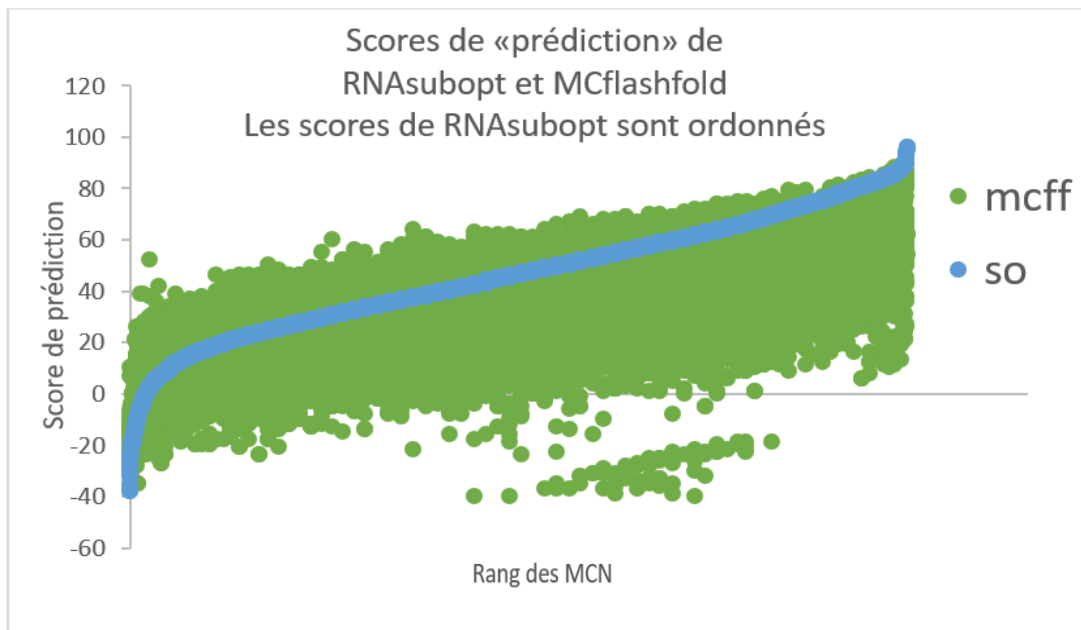
Évaluation du score de prédiction

Score de prédictions de la SS de de la *MFE* dans l'ensemble non filtré

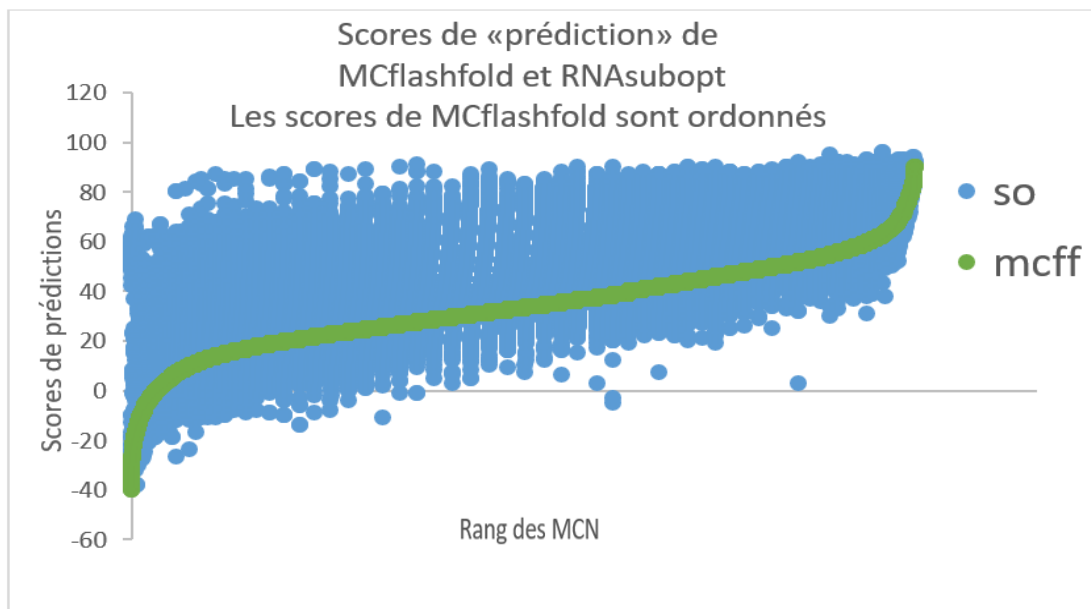
Les ARN de la figure ci-dessous sont classés par leur différence de score prédit entre les deux logiciels. On observe que les prédictions de MCFlashfold sont en moyenne moins précises que celle de RNAsubopt de 12.5 points. Cela signifie qu'en moyenne pour un ARN les prédictions de 12.5 nt. sont soit ratés par MCFlashfold pendant que RNAsubopt ne se prononce pas, soit le contraire, soit un mélange des deux. Puisqu'avec ce nombre de données, il est rare que le logiciel ne se prononce pas, il est plus probable qu'en moyenne six prédictions de MCFlashfold soient ratées pendant que RNAsubopt en réussit six de plus. Les deux figures de la page suivante montrent les scores absolus. Le coefficient de corrélation de Pearson est relativement élevé, il est légèrement au-dessus de 0.76. Cela indique que certains ARN sont plus difficiles à prédire que d'autres peu importe la méthode.



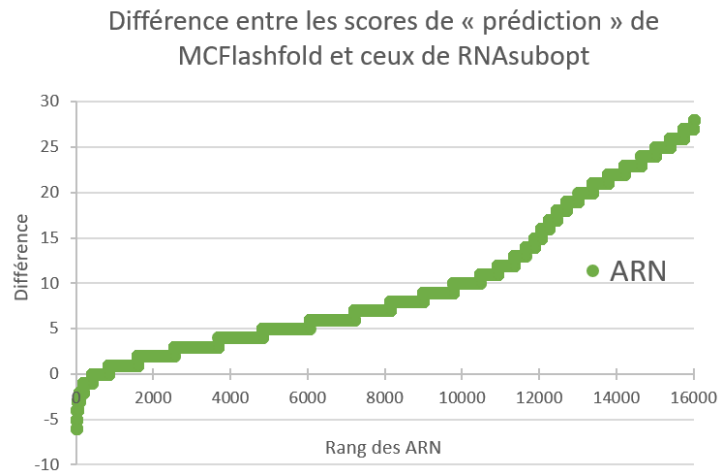
La différence de score de prédiction entre MCFlashfold et RNAsubopt est en moyenne de 12.5 points en faveur de RNAsubopt. Les pires ARN ont une forte concentration en adénine ou ont un score moyen élevé



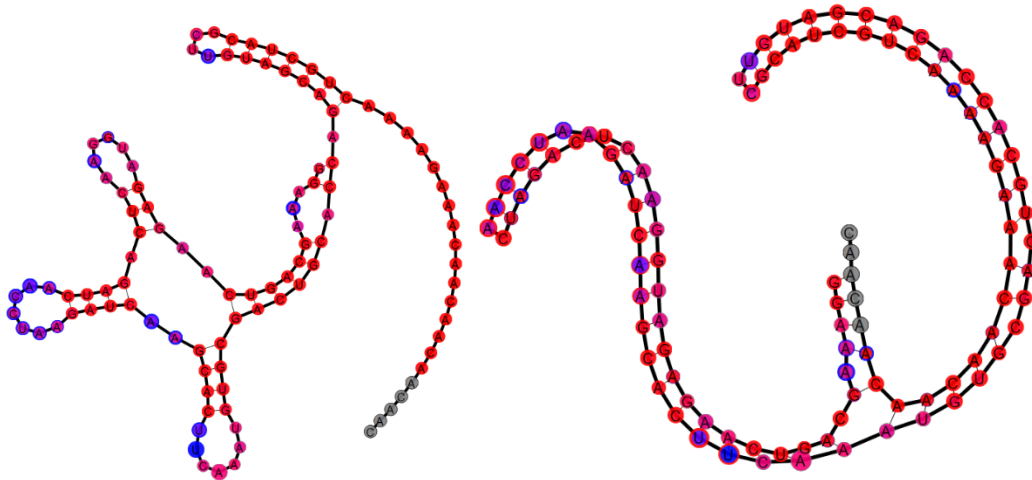
Le coefficient de corrélation de Pearson entre les scores de prédiction des deux logiciels est de 0.763083. Ceci indique une bonne corrélation entre les deux logiciels. Le score de MCFlashfold est en vert et celui de RNAsubopt en en bleu.



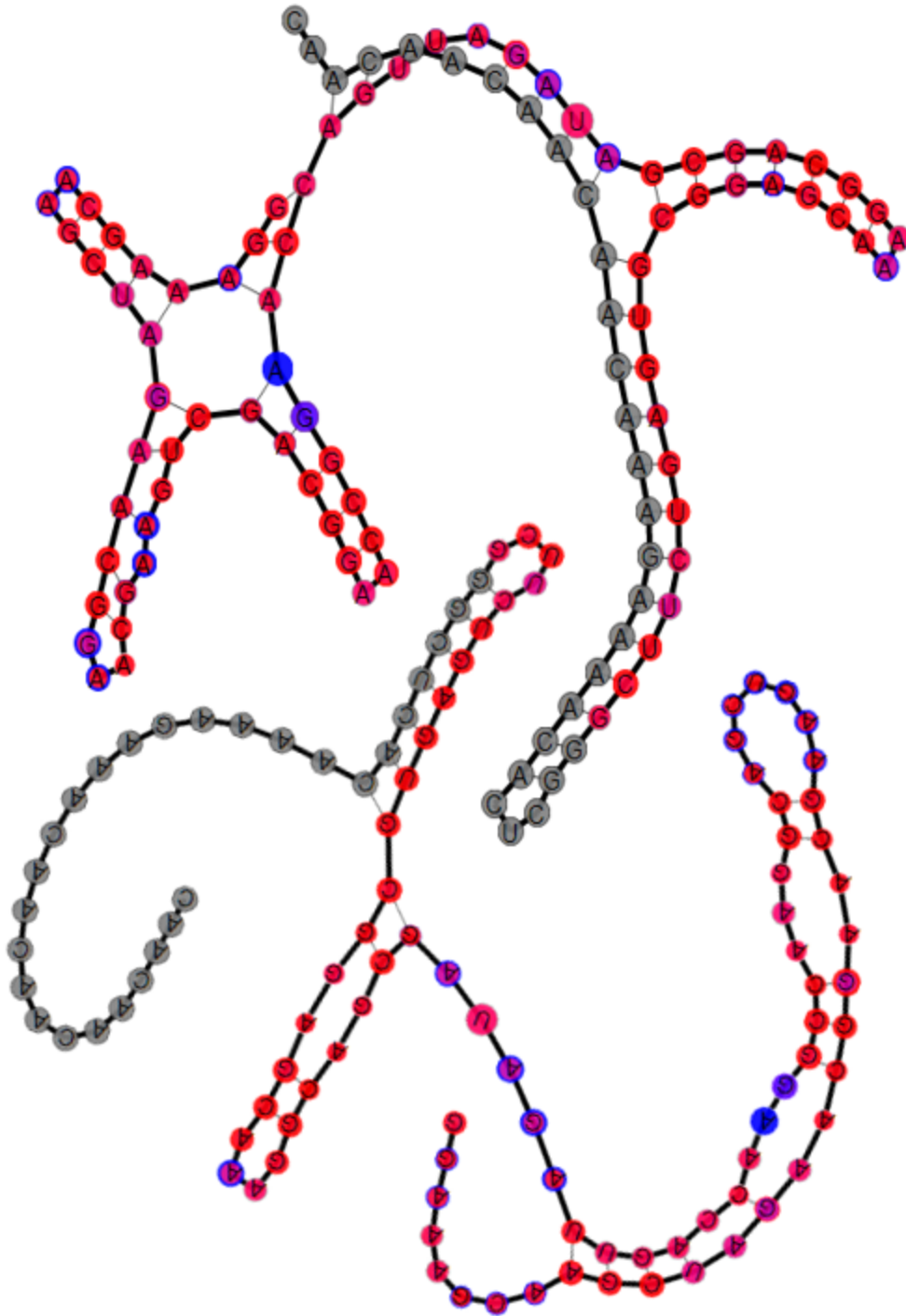
Cette figure est la figure du haut inversée. Ici, ce sont les données de MCFlashfold qui sont ordonnées. La corrélation est visible par le nombre presque inexistant d'ARN de RNAsubopt (bleu) en bas à droite



Après filtration des données, la moyenne des différences passe à 10.5 soit un gain de 2 points en faveur de MCFlashfold. L'étendue des différences s'est amoindrie aussi. À la défense de MCFlashfold, les données ont été calibrées à partir d'une structure en épingle prédite par un logiciel de la même suite que RNAsubopt, ViennaPackage [40].



Correspondance entre les valeurs de sondage chimique et les nt. d'un ARN repliés par RNAsubopt à gauche et MCFlashfold à droite. Le score de prédiction est de 46 contre 8 pour RNAsubopt. Lorsqu'on utilise le score pondéré, RNAsubopt obtient 66.62 et MCFlashfold 28.21.

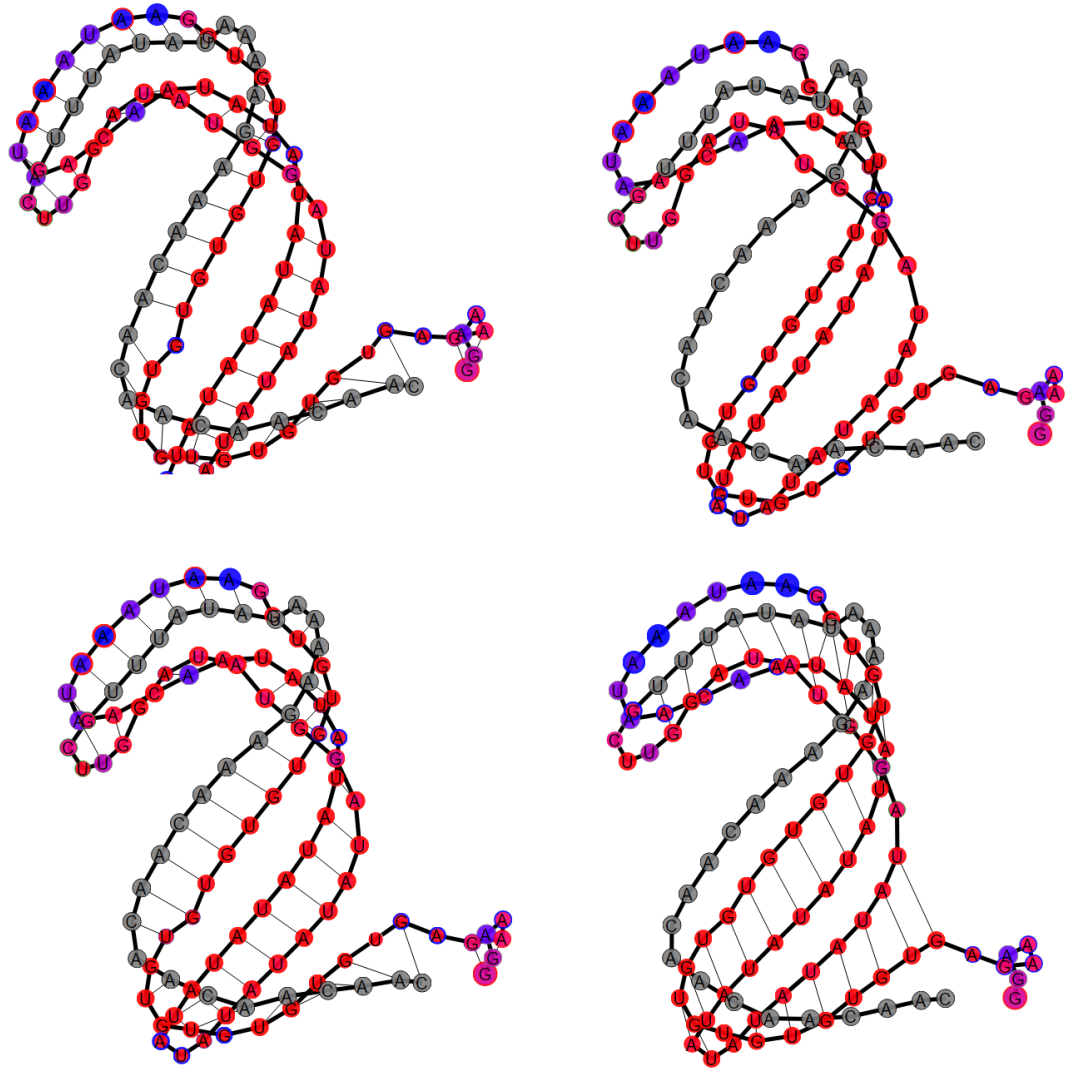


ARN ayant la différence de score la plus avantageuse pour MCflashfold. La différence entre les scores de prédiction pondérés est de 6.35. Les scores sont de 22.52 pour MCFlashfold et de 16.17 pour RNAsubopt. L'ARN du haut a été replié par MCFlashfold et celui du bas par RNAsubopt.

Ce tableau présente les cycles les plus souvent retrouvés dans la base de données. Le ratio donne une idée de la pureté du cycle. Un ratio de 1 signifie que tous les nt. du cycle sont réactifs, tandis qu'un ratio de -1 signifie que tous les nt. de la S-S sont peu réactifs.

Identifiant du cycle	Nombre de nt. total so	Différence so	Nombre de nt. total mcff	Différence mcff	ratio So	ratio MCff
2_2-GC-GC_pos_0	125470	-112116	109185	-97280	-0.89	-0.89
2_2-GC-GC_pos_1	124541	-112696	108655	-97635	-0.9	-0.89
2_2-UA-UA_pos_1	112327	-86408	95831	-74733	-0.76	-0.77
2_2-UA-UA_pos_0	111662	-92867	95730	-79500	-0.83	-0.83
2_2-AU-AU_pos_0	106443	-76420	97478	-66774	-0.71	-0.68
2_2-AU-AU_pos_1	103962	-88838	96290	-77878	-0.85	-0.8
2_2-GA-UC_pos_1	102951	-73121	92271	-64809	-0.71	-0.7
2_2-UC-GA_pos_0	102619	-93206	97151	-86758	-0.9	-0.89
2_2-GA-UC_pos_0	102365	-79364	92251	-69446	-0.77	-0.75
<i>contrôle visuel</i> →					-0.5	-1

Portrait d'un ARN : Représentation en deux dimensions de quatre des multiples conformations possibles de cet ARN.



Séquence :

GGAAAGAGUGUGUGUGUGUGUGUGUGUGUGUGGAAUAAAUACAAUAUA
UAUAUAUAUAUAUAUAUAUAGGUUAUAGGUUCGUUAUAUAAAAGAAA
CAACAACAACAAC

Le graphe des transitions d'un ARN

Les deux algorithmes donnent en sortie des SS classées en ordre selon leur énergie libre, ordonnée de la plus basse vers la plus élevée. Les SS prédites sont plus ou moins différentes les unes des autres selon plusieurs aspects : le nombre d'hélices, la composition en bases des jonctions, le nombre et l'identité des paires de bases, etc. La plus grande différence est que *MCFlashfold* permet les paires de bases non canoniques tandis que *RNAsubopt* permet un type de paire de base non canonique : la paire GU. Une valeur de distance ou de similarité peut être calculée entre chacune des SS, ce qui permet par la suite une représentation sous forme de réseau. Ce réseau est nommé le graphe des transitions.

Dans sa forme la plus simple, le calcul de la distance entre deux structures se fait en dénombrant le nombre de paires de bases communes [41]. Plus ce nombre est élevé, plus les deux SS sont semblables. Une façon plus stricte en général de relier les SS est de considérer le changement d'une seule paire de base, Dr Paul Dallaire en discute dans sa thèse [42] et fait référence à un article de Dr Christoph Flamm [43].

L'ensemble des structures (nœuds) et des liens entre les SS semblables se nomme : le graphe des transitions, puisque l'on fait l'hypothèse qu'une structure peut transiter en une autre lorsqu'elles sont reliées dans le graphe. À ma connaissance, les règles de ces transitions n'ont pas été testées de façon approfondie en laboratoire. Actuellement, ce réseau nous permet de comparer la diversité des SS prédites entre deux ARN différents et il nous donne une idée du paysage énergétique de la molécule.

En se repliant, un ARN emprunte des chemins dans ce réseau passant d'une forme de haute énergie à une forme de basse énergie. En clair, l'ARN tentera de minimiser son énergie en créant des structures stables telles que des paires de bases ou des boucles de basses énergies. L'ensemble des chemins menant vers un minimum local est appelé un puits d'énergie et l'une des causes du mauvais repliement d'un ARN est la présence de plusieurs puits d'énergie semblables piégeant l'ARN dans un minimum local.

Une explication non validée de la difficulté à prédire la réactivité est l'absence de considération de la structure tertiaire, c'est-à-dire les interactions à longue distance.

Base de données RMDB

La « RNA mapping data base » est une base de données des expériences de sondages chimiques de l'ARN. Elle contient plusieurs types d'expériences. Certaines sont faites sur des ARN provenant de cellules vivantes et d'autres sont faites sur des ARN synthétisés en laboratoire. Elle fournit les séquences d'ARN avec une valeur associée aux nt. donnant la mesure de leurs réactivités. Les données utilisées dans ce mémoire sont normalisées sur une séquence formant une boucle ajoutée en 3'. Cette boucle a une réactivité connue et constante.

Lorsque la valeur de réactivité chimique est basse, cela signifie une absence de réactivité et une valeur haute signifie une réactivité élevée.

J'ai classé les nt. en deux groupes, les nt. qui réagissent peu, les « Low » et ceux réagissant fortement les « Hi ». Il y a une zone de réactivité qui n'est pas considérée, elle correspond aux nt. qui sont entre le « Low » et le « Hi ».

La RMDB contient 3 catégories d'expériences :

1. Des expériences publiées ou provenant du laboratoire du Dr Das.
2. Des expériences utiles aux puzzles sur l'ARN (RNA puzzles).
3. Des expériences reliées à ETERNA.

Mes analyses ont été faites sur les données reliées à ETERNA.

ETERNA

Eterna est un jeu vidéo scientifique. Il propose aux joueurs des casse-têtes ayant pour thème l'ARN. Un des jeux consiste à trouver une séquence qui se replie en une SS donnée, c'est l'inverse du problème de repliement résolu par RNAsubopt et MCFlashfold.

Cloud Labs

Lorsqu'un joueur expérimenté a trouvé une séquence se repliant en la structure demandée, il peut la soumettre au laboratoire pour confirmer sa découverte. Le laboratoire synthétise l'ARN et établit son profil par sondage chimique.

Fonction d'évaluation des séquences soumises

Dans ce mémoire, pour évaluer les SS quant à leur degré de similarité avec les données de SHAPE, je me suis inspiré de la fonction de score d' ETERNA. Dans ce jeu, les séquences soumises sont évaluées en fonction de la réactivité de leurs bases. Un seuil est déterminé et si le nucléotide est censé être païré dans la SS cible, mais qu'il réagit à l'agent modificateur, un point est retiré du total. Dans le cas contraire, un nt. non païré doit ne pas réagir du tout pour qu'un point soit retiré. Cette évaluation est très « gentille » pour le joueur, puisqu'un nucléotide tombant dans la zone grise n'entraîne pas de pénalité et que les valeurs de seuil sont optimisées pour maximiser le score du joueur. Ceci est dû à l'incertitude de la vraie signification de la réactivité chimique d'un nucléotide. Cette valeur est nommée : « Structure mapping score » [44]. Je me suis inspiré de ce score pour distinguer les nt. réactifs des nt. peu réactifs. La seule différence est que les seuils sont fixes.

Dans ce mémoire, vous constaterez que le concept païré ou non païré peut être précisé pour mieux évaluer la réactivité ou la non-réactivité d'un nucléotide

Mapseeker

Mapseeker est le logiciel faisant la correspondance entre les séquences déterminées lors de l'étape de séquençage et la valeur de réactivité pour chaque nucléotide de la séquence. Son fonctionnement est décrit dans [33]

Prise en considération des décrochements naturels de la polymérase

Comme mentionné brièvement plus haut, un contrôle négatif est effectué pour chaque ARN sondé. Ceci permet de distinguer les arrêts de transcriptions non provoqués par l'agent modificateur. Par exemple une SS très stable peut empêcher la rétrotranscriptase de poursuivre sa transcription.

Normalisation sur un segment connu

L'ajout d'une tige-boucle permet de mesurer chaque réaction de façon extrêmement précise. Dans le passé, comparer les expériences faites sur différents ARN était une tâche presque impossible. Il est maintenant beaucoup plus facile de le faire grâce à cet ajout.

Le biais de ligation

Certaines séquences ont moins de chance de se faire transcrire par la polymérase, créant ainsi un biais. Lors de la calibration des méthodes de détermination de la réactivité chimique des ARN par séquençage, les auteurs de [33] ont remarqué que les données de réactivités normalisées adjacentes à la tige-boucle « GAGUA » sont plus élevées que ce qu'ils avaient découvert avec les expériences faites par électrophorèse capillaire. Ils ont déterminé un facteur pour corriger cette différence.

Rapport « signal sur bruit » (*Signal to noise ratio*)

Cette valeur est calculée en divisant la moyenne des valeurs de réactivité par la moyenne des erreurs de tous les nt. d'un ARN [45]. Pour plus d'information sur l'erreur, on peut se référer à l'article de Matthew G. Seetin [34]. On y apprend entre autres que l'erreur est basée sur la loi de poisson.

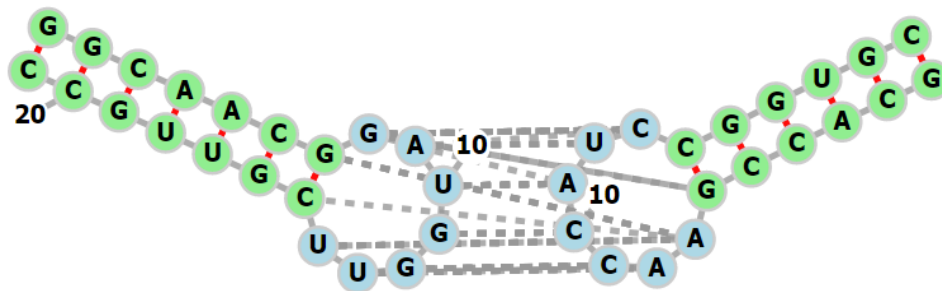
Outils de visualisation des structures secondaires

VARNA

La base de données RMDB utilise un outil nommé VARNA pour représenter la SS et les données de sondage. Cet outil est écrit en java et le code source est offert gratuitement sous la licence GNU GLP. Trois limitations m’ont poussé à créer m’ont propre outil. Premièrement, Java contrairement à JavaScript n’est pas interprété directement dans le navigateur, il doit être compilé par une machine virtuelle non native et gérée à l’intérieur d’un « applet ». La deuxième raison est qu’il ne permet pas de visualiser le graphe des transitions des ARN. Finalement, il ne permet pas de mettre l’accent sur un nucléotide en particulier [46].

FORNA

FORNA est un logiciel de représentation de la SS écrit en JavaScript et utilisant la même technologie que RNA Dynamic Viewer (RDV) (mon logiciel), soit d3js. FORNA est offert gratuitement sur *github* (un répertoire en ligne de code informatique basé sur git un logiciel de gestion des versions). La seule raison pour laquelle, je ne l’ai pas utilisé c’est qu’il n’existait pas lors du début de la conception de RDV, en 2014 [47].



Représentation en deux dimensions de deux ARN provenant de la « *protein data bank* » (PDB) par le logiciel FORNA. Les structures se nomment « *RNA LOOP-LOOP COMPLEX* ». Leur identifiant est : « 1BJ2 ».



Représentation en trois dimensions des ARN de la figure de la page précédente. Les pointillés mauves sont des ponts hydrogènes. La couleur des nt. est reliée à leur position allant du bleu (5') au rouge (3')

Azure

Azure est une plateforme développée par l'équipe de Microsoft ®. Elle est fournie gratuitement aux étudiants de l'Université de Montréal en version d'essais. Elle permet de faire de l'apprentissage machine de façon intuitive et graphique. Plusieurs algorithmes sont offerts et ils sont tous paramétrables. J'ai utilisé la plupart d'entre eux et j'ai comparé leur efficacité dans la tâche de prédire la réactivité d'un nucléotide sur différentes entrées. Un réglage automatique des hyperparamètres est possible, ce qui permet une utilisation par un scientifique non expert en apprentissage automatisé. De plus, c'est une très bonne façon de s'instruire sur le sujet tout en testant les algorithmes.

Classifieurs

Les classifieurs sont des algorithmes qui prennent en entrée des éléments ayant tous les mêmes classes de caractéristiques et qui produit des catégories classant ces éléments. Les classifieurs suivants ont été utilisés pour prédire la classe de réactivité des nt. (basse ou haute).

Réseau de neurones

Un réseau de neurones est une classe d'algorithme d'apprentissage machine complexe. En général, les réseaux de neurones sont composés de nœuds et de liens entre les nœuds ayant chacun un poids. Ils sont faits de plusieurs couches de nœuds. Il y a la couche d'entrée où chaque nœud correspond à une dimension du problème à résoudre. Ensuite, il y a une ou plusieurs couches cachées. Dans un réseau de neurones classique, chaque nœud de la première couche cachée est relié à tous les nœuds d'entrées.

Régression logistique

La régression logistique est un modèle d'apprentissage machine populaire dans lequel on tente de maximiser la vraisemblance d'un modèle logistique de la forme :

$$P(D=1|X_1, X_2, \dots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

« D » est la variable à prédire et les « X » sont les caractéristiques données en entrée. Les paramètres à déterminer sont « α » et les « β ». Plus le « β » d'une variable est élevé plus l'importance de la variable est grande dans le modèle. Lorsque la variable est catégorique, elle est transformée en variable binaire au début de l'algorithme [48].

Machine à vecteurs de support.

Ce sont des algorithmes qui tentent de trouver un hyperplan qui sépare les deux classes à prédire. Pour avoir plus de détail sur les différents types de machine à vecteur de support ou « *support vector machine* » (SVM) en anglais, je vous conseille de lire [49].

Arbres de décisions

Les arbres de décisions sont des structures de prises de décisions extrêmement simples. Comme leur nom l'indique, leur structure est en forme d'arbre. À chaque

nœud, une décision est prise, éliminant du même coup les ramifications non explorées. Lorsqu'il est binaire, seulement deux branches sont accessibles par nœud. Au bout de chaque chemin, constitué de nœuds et de branches, se trouve la réponse. Pour entraîner un arbre de décision, on génère des arbres aléatoirement et l'on compare leur succès par des métriques tel que ceux expliqués plus loin, dans la section : « Métriques de mesure de la performance des algorithmes d'apprentissage machine ». On peut aussi utiliser un algorithme génétique pour combiner les arbres qui performant bien. Pour plus de détail référez-vous à [50]

Arbre de décision « boosté » ou « *Boosted Decision Tree* »

Cette méthode de classification s'appuie sur les arbres de décision. De façon itérative, on tente de diminuer l'erreur des arbres de décision modélisés au paravent. Pour plus de détails, veuillez-vous référer à [51].

« *Bayes point machine* »

Cette technique de classification repose sur la probabilité d'observer une caractéristique étant donné une autre caractéristique. On ajoute à cela un biais pour maximiser l'algorithme. Pour plus de détails, veuillez-vous référer à [52].

Métriques couramment utilisées

Dans cette section plusieurs métriques sont expliquées. Grossièrement, une métrique permet de mesurer un phénomène. Les métriques de corrélation prennent en entrée deux vecteurs de la même taille, tandis que les métriques de mesure de performance des algorithmes d'apprentissage machine présentée ici prennent en entrée un classifieur binaire tels que ceux présentés plus haut.

Corrélation de Pearson (corrélation linéaire)

Cette métrique permet d'établir une corrélation entre deux vecteurs de la même longueur. Chaque élément d'un vecteur doit être relié à l'élément de même indice de

l'autre vecteur. Une valeur de 1 signifie une corrélation parfaite et une valeur de -1 signifie une corrélation inversée. Une valeur de 0 signifie : « pas de corrélation » du tout. D'autres métriques de mesure de la corrélation existent, par exemple la corrélation de Spearman et la covariance.

Métriques de mesure de la performance des algorithmes d'apprentissage machine

Les vrais positifs

Lorsqu'on prédit des éléments pouvant se trouver dans deux classes distinctes, les positifs et les négatifs, les vrais positifs sont de vraies (bonnes) prédictions d'un élément positif.

Les faux positifs

Tous comme les vrais positifs, les faux positifs sont un jugement sur la prédiction d'un élément pouvant être soit positif ou négatif. Les faux positifs sont donc de fausses (mauvaises) prédictions ayant été prédites dans la classe des positifs. Les prédictions sont faussement positives.

Les vrais négatifs et les faux négatifs

Il suffit de remplacer le terme positif par le terme négatif dans les deux définitions précédentes pour avoir la définition des vrais et faux négatifs.

Dans le reste du mémoire, il sera question de vrais « Hi », faux « Hi », vrais « Low » et faux « Low », cela fait référence à la prédiction des classes discrètes de réactivité chimique des nt.. Les seuils sont de 0,5 et 1. (« Hi » > 1 & « Low » < 0,5)

Le taux de vrais positifs (« *True positive rate* »)

Le taux de vrais positifs est le nombre de vrais positifs divisé par le nombre d'éléments positifs total. L'axe vertical de la courbe ROC est le taux de vrais positifs.

Le taux de faux positifs (« *False positive rate* »)

Le taux de faux positifs est le nombre de faux positifs divisé par le nombre total d'éléments négatifs. Autrement dit, c'est le nombre d'éléments prédits positifs, mais qui sont en réalité négatifs divisé par le nombre total d'éléments négatif. L'axe des abscisses (horizontale) du plan de la courbe ROC est le taux de faux positifs. En général, moins la sensibilité ou le rappel est grand, moins le taux de faux positifs est grand, c'est ce que la courbe « ROC » illustre. Ceci est vrai pour les algorithmes tel que les classifieurs bayésiens naïfs puisqu'ils donnent une probabilité d'appartenance à une classe [41]. Plus on augmente le seuil de risque que prend l'algorithme, plus le nombre total de vrais positifs augmente, mais de moins en moins rapidement. L'augmentation du risque que prend l'algorithme entraîne aussi le nombre de faux positifs à la hausse. Un algorithme risqué prédit plus d'éléments dans la classe des positifs qu'un algorithme peu risqué. Un bon algorithme a un rapport élevé entre le taux de vrais positifs et le taux de faux positifs.

La précision

Cette valeur s'étend aussi de 0 à 1. La précision est le nombre d'éléments bien prédit divisé par l'ensemble des éléments prédits dans cette classe. Conceptuellement, moins l'algorithme prend de risque plus la précision devrait être élevée (moins de fausses prédictions). Cependant, le nombre total de vrais positifs est par le fait même diminué. Pour bien comprendre cette notion, on peut se placer dans le contexte d'un test diagnostique. Un jugement positif signifie que le patient a la maladie. Pour avoir un test précis, donc avec un taux de faux positifs bas, il faut diminuer le nombre de déclarations positives totales et se restreindre aux patients ayant des signes évidents. Dans le cas de la prédiction de la réactivité chimique des nt., une précision élevée déclare comme réactifs une petite fraction des nt. qui le sont, mais se trompe rarement. Pour atteindre une plus grande précision, on doit donc simplement prendre moins de risques et prédire moins d'éléments de la classe d'intérêt.

Le rappel ou sensibilité

C'est le nombre d'éléments bien prédit d'une classe divisée par le total des éléments de la classe. Un ratio de 1 signifie que tous les éléments de la classe ont été trouvés et bien prédits. Dans l'exemple d'un algorithme avec un seuil de risque au-dessus duquel il se prononce, il y a un compromis à faire entre le rappel et la précision. La solution la plus simple pour avoir un rappel de 1 est de prédire tous les éléments comme étant positifs. Cependant, la précision risque d'être moins grande.

Score F1

Cette valeur considère le rappel et la précision dans son calcul. Plus précisément, le score F1 et la moyenne harmonique de la précision et du rappel. Plus la valeur est près de 1, meilleur est l'algorithme [53].

Table de contingence

Cette table est une matrice $n \times n$, « n » étant le nombre de classes possibles dans laquelle le nombre de vraies « classes i » et de fausses « classes i » est indiqué, i étant le nom des classes, voir tableau VIII. La table de contingence est une généralisation de la matrice de confusion.

Courbe ROC

Lorsqu'un algorithme donne une probabilité d'appartenance à une classe donnée, un seuil est nécessaire pour départager les prédictions. Si le seuil est un nombre réel et qu'il peut être ajusté, on peut générer « une infinité » de table de contingence. La courbe ROC permet de voir un portrait d'ensemble de ces matrices. Chaque point de la courbe représente le taux de vrai positif en fonction du taux de faux positifs.

Autres logiciels utilisés

Mongodb

MongoDb est une base de données non relationnelle (BDNR). À la différence des bases de données relationnelles (BDR), les BDNR peuvent contenir des structures de données en forme d'arbre. Chaque document ajouté à la base de données à un identifiant unique. Comme dans la plupart des BDR, quatre commandes de base permettent d'interagir avec la base de données : « insert, remove, update et find ».

On doit créer la base de données en lui donnant un nom. À l'intérieur de celle-ci se trouvent des collections, c'est dans ces contenants que sont placées les données. Chaque collection est indépendante. Pour accélérer la recherche de la valeur d'un des champs, on « index » ce champ avec la commande *createIndex*. MongoDB m'a permis de compiler les S-S comme vous pouvez le lire dans le chapitre 1.

Node.js

Node.js est un serveur web. Contrairement au serveur « Apache », il n'interprète pas le langage PHP, il interprète le JavaScript, ce qui est pratique dans la mesure où JavaScript est le langage interprété par les navigateurs. Dans mon cas, JavaScript a été essentiel pour la réalisation des interfaces interactive et dynamique. La combinaison de Node.js et de MongoDB est facile à mettre en place. RDV utilise ces deux technologies.

Formats de fichiers

RNASS produit en sortie des fichiers JSON qui peuvent être convertis en fichier CSV pour faciliter leur étude ultérieure.

JSON (JavaScript object notation)

Ce format est largement utilisé. Il permet de conserver des données dans une structure en forme d'arbre. Il peut être lu par des éditeurs de texte directement. Il est compatible avec la base de données MongoDB.

CSV (comma separated value)

Ce format est très simple. Tout comme le format JSON, il peut être lu directement par un éditeur de texte. Il range les données sous forme de tables, c'est-à-dire une liste de rangées avec un nombre de champs définis formant des colonnes. Les rangées sont séparées par des sauts de ligne et les colonnes par des virgules. Il est compatible avec logiciel Excel de Microsoft. Les fichiers TSV (*tab separated value*) diffèrent des fichiers CSV par le remplacement des virgules par des tabulations.

Le format RDAT

Les données de la RMDB sont disponibles en deux formats : le format ISATAB (non utilisé) et le format RDAT. Les spécifications de ces formats sont disponibles sur le site web de la RMDB [54]. Les fichiers RDAT sont générés par MapSeeker [34]. Un outil pour les interpréter est mis à notre disposition, il se nomme RDATKit. Cette librairie est écrite en *Python* et est particulièrement complexe. Un fichier RDAT correspond à une expérience distincte pouvant contenir des milliers de séquences sondées. Chaque fichier a une liste de « *constructs* », qui contient la majorité du temps un seul élément. Cet élément a un champ « *data* » contenant l'ensemble des données de réactivités, les données relatives aux erreurs entre les répliques, les séquences et le « *signal sur bruit* ». Les séquences sont parfois plus longues que les vecteurs de réactivités et d'erreurs. Pour phaser les deux vecteurs, un champ nommé « *offset* » accompagne chaque séquence, les nt. avant cet indice n'ont pas de valeur de réactivité. Les conditions dans lesquels l'expérience a été menée, tel que la température et la concentration des espèces chimiques, sont disponibles dans la variable annotation de

chaque séquence. De plus, c'est dans ce fichier qu'on retrouve le traitement numérique qui a été fait sur les données.

La figure de la page suivante montre la distribution des cycles simples quant à leur fréquence d'apparition. La distribution des cycles composés est semblable. Le nombre de cycles simples possibles est très grand. Le calcul du nombre de cycles simples de 6 nt. et moins le montre clairement.

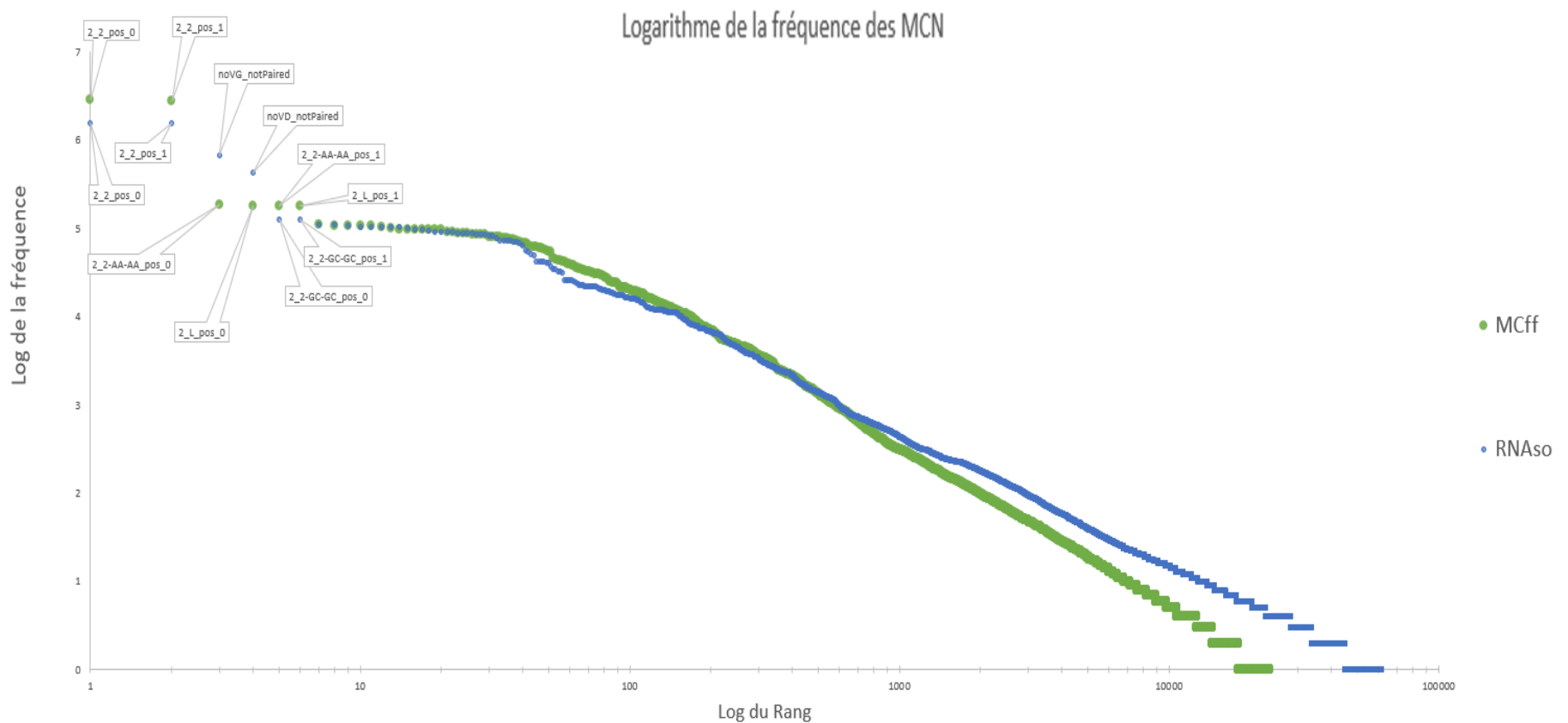
$$2 \times 4^4 (\text{«2_2»}) + 5 \times 4^5 (\text{«3_2»}) + 3 \times 4^6 (\text{«3_3»}) + 6 \times 4^6 (\text{«2_4» et «4_2»}) = 42\,496$$

La symétrie des cycles ayant le même nombre de nt. de chaque côté a été prise en compte. La combinaison de deux cycles multiplie les possibilités. Le résultat des trois positions possibles des nt. dans le cycle «3_4» multiplié par 4^7 (nombre qui représente le choix des 4 nt. à chacune des positions) donne déjà : 49 152. Il faut noter que ce chiffre dépasse le nombre réel de possibilités puisque ce ne sont pas tous les nt. qui peuvent former des paires de base pour le logiciel RNAsubopt.

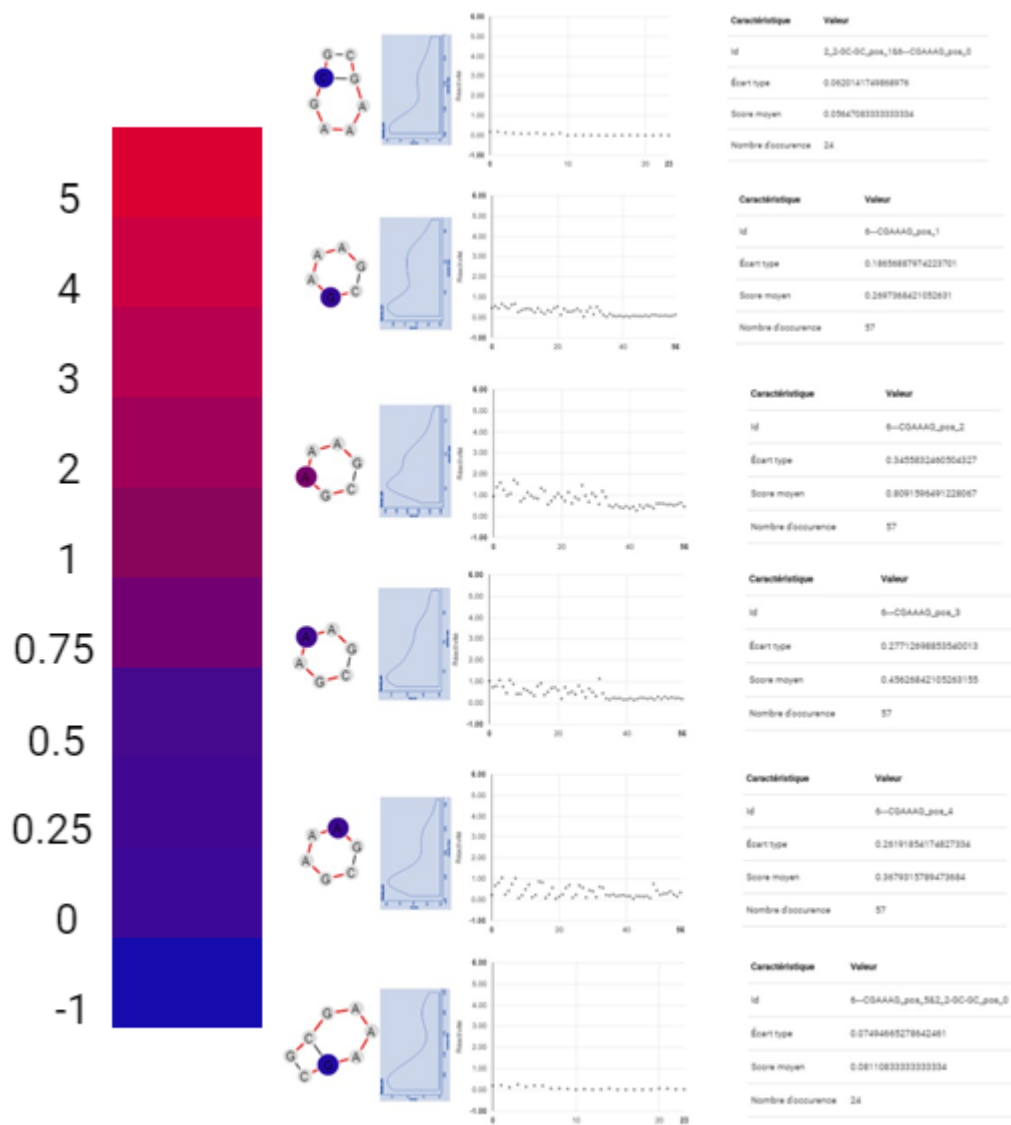
Le logiciel MCFlashfold quant à lui à l'avantage de former des plus petits cycles ce qui limite le nombre de S-S en pratique.

Quoi qu'il en soit, le nombre de S-S est élevé.

Les deux figures suivantes montrent l'occurrence des S-S (anciennement nommées MCN) pour les deux logiciels de repliement des ARN. Une transformation logarithmique a été appliquée sur les deux échelles. Chaque S-S d'un logiciel est mis en correspondance avec son homologue dans l'autre logiciel.

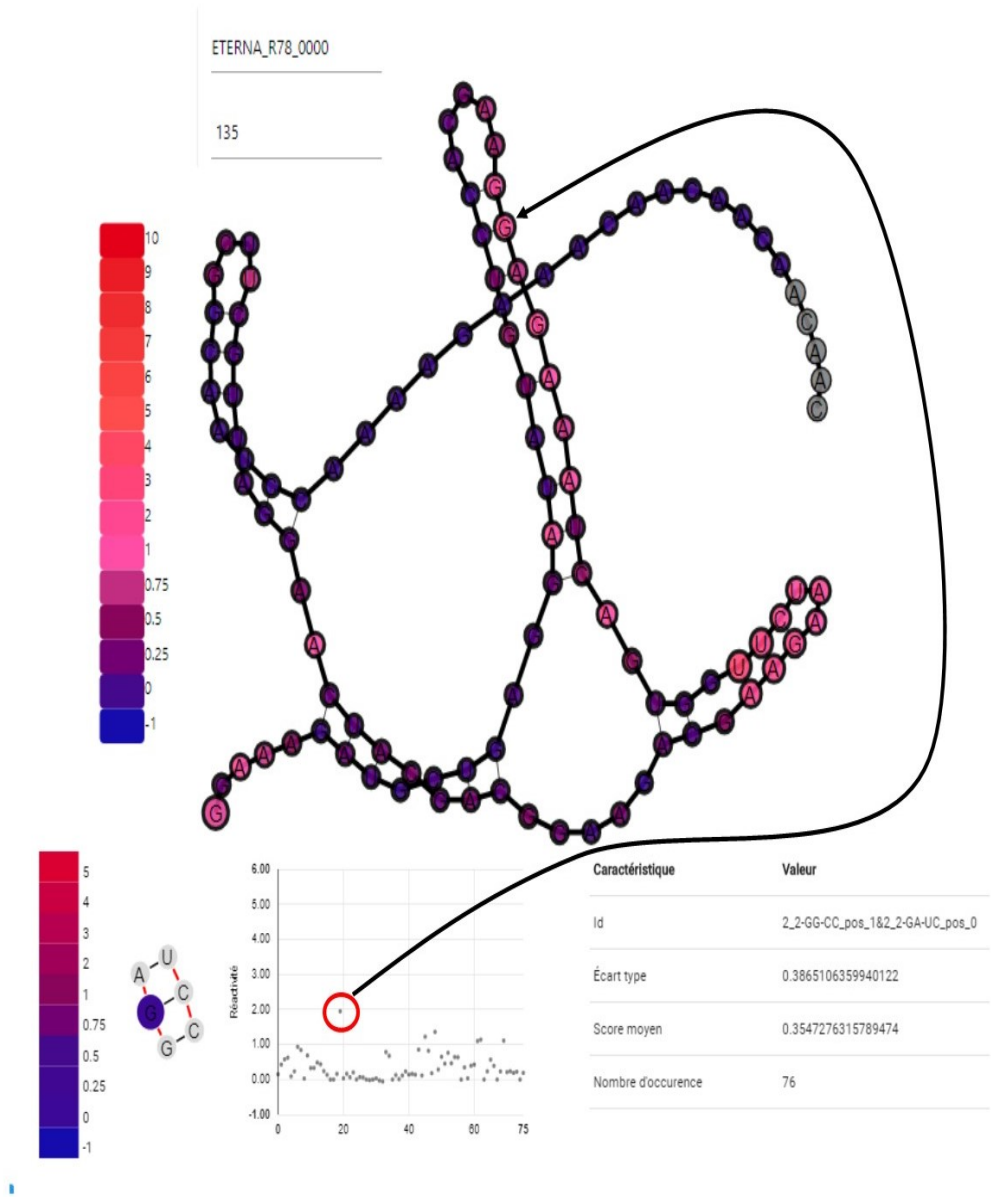


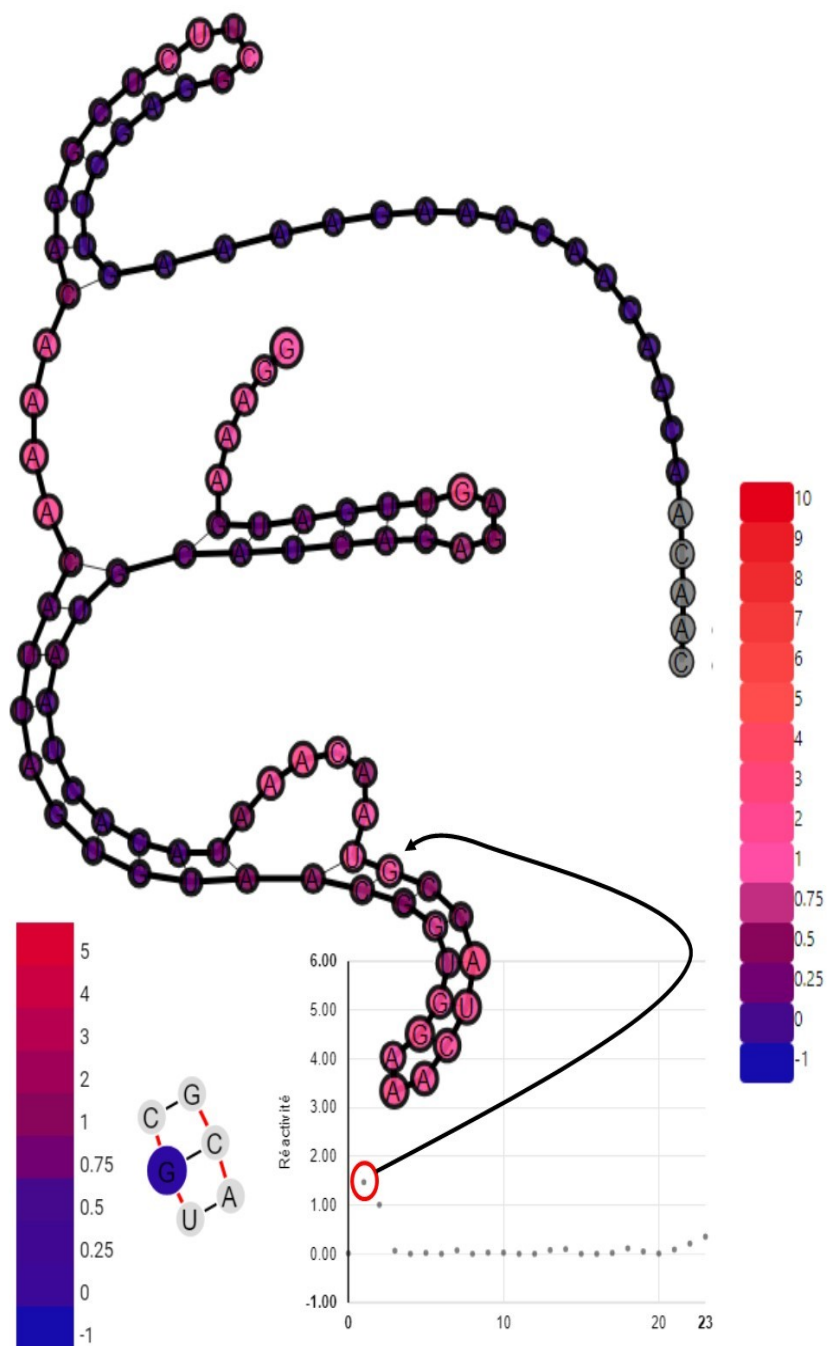
Distribution des cycles simple en fonction de leur fréquence. Une transformation logarithmique de base 10 a été appliquée sur les deux axes pour mieux voir les cycles simples les plus fréquents. Chaque logiciel est indépendant par rapport à l'ordre de ses cycles simples. Les cycles simples du logiciel de RNAsubopt sont représentés par des points bleus et les cycles simples de MCFlashfold sont représentés par des points verts. Certains cycles simples caractérisent plusieurs millions nt. dans la base de données complète de RMDB. Pour être considéré, le nucléotide doit être dans la conformation du cycle simple dans plus de 20% des SS sous-optimales. On trouve dans les boîtes des cycles simples les plus fréquents leur identifiant.

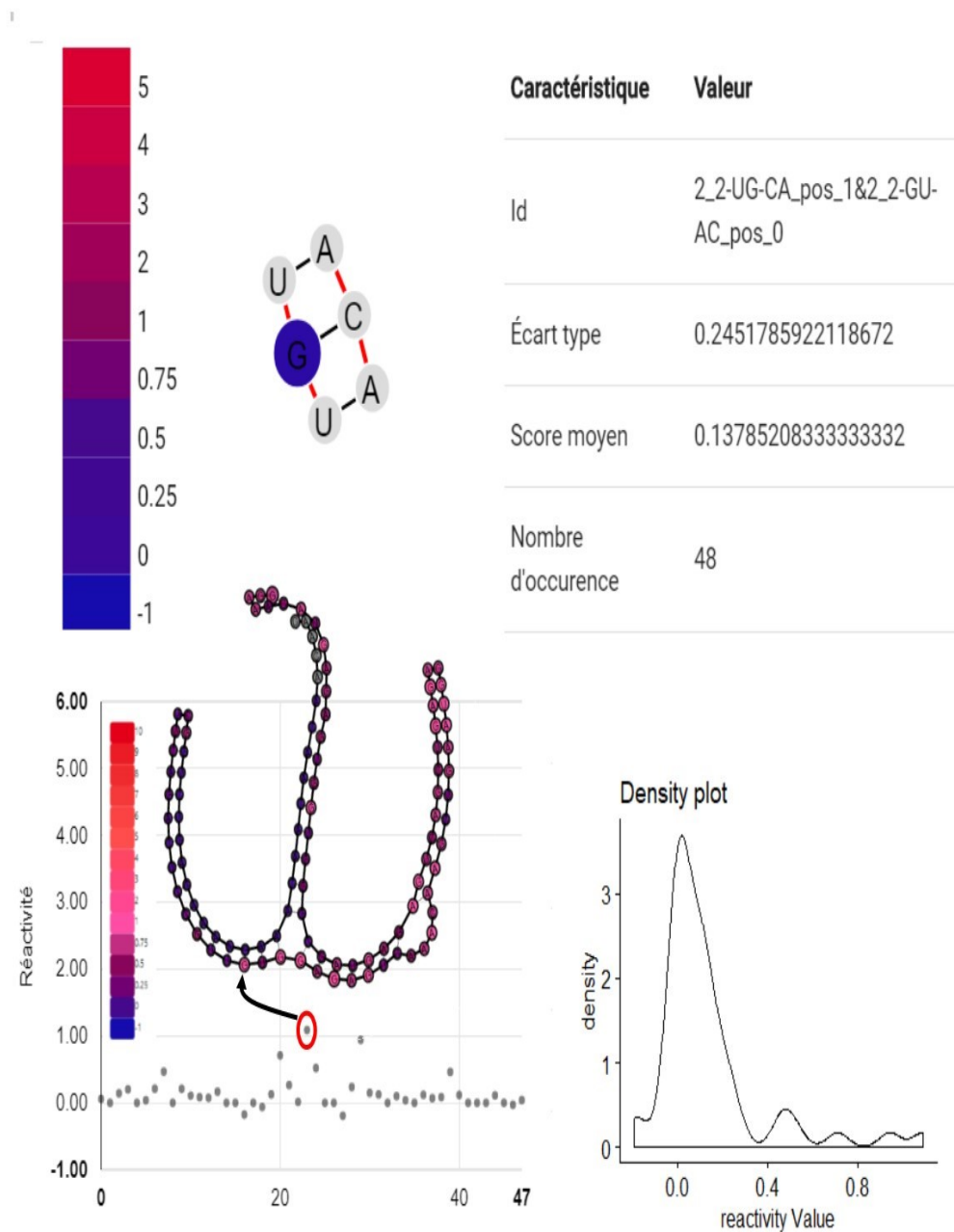


Cette figure représente la réactivité d'une boucle « CGAAAG » nucléotide par nucléotide (nt.). On voit que chaque nt. est plus ou moins réactif et que le nt. ayant la moyenne la plus élevée est l'adénine adjacente à la guanine.

Les figure des trois prochaines pages illustrent une utilisation de RDV. Elles montrent qu'il est possible d'étudier en détail les nt. ayant des scores de réactivités anormaux.







Le « *density plot* » représente la distribution des scores de réactivité d'une S-S.